Dealing with Sparse Rater Scoring of Constructed Responses within a Framework of a Latent Class Signal Detection Model

Sunhee Kim

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee of
the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Dealing with Sparse Rater Scoring
of Constructed Responses within a Framework
of a Latent Class Signal Detection Model

Sunhee Kim

In many assessment situations that use a constructed-response (CR) item, an examinee's response is evaluated by only one rater, which is called a *single rater design*. For example, in a classroom assessment practice, only one teacher grades each student's performance. While single rater designs are the most cost-effective method among all rater designs, the lack of a second rater causes difficulties with respect to how the scores should be used and evaluated. For example, one cannot assess rater reliability or rater effects when there is only one rater.

The present study explores possible solutions for the issues that arise in sparse rater designs within the context of a latent class version of signal detection theory (LC-SDT) that has been previously used for rater scoring. This approach provides a model for rater cognition in CR scoring (DeCarlo, 2005; 2008; 2010) and offers measures of rater reliability and various rater effects. The following potential solutions to rater sparseness were examined: 1) the use of parameter restrictions to yield an identified model, 2) the use of informative priors in a Bayesian approach, and 3) the use of back readings (e.g., partially available 2nd rater observations), which are available in some large scale assessments. Simulations and analyses of real-world data are conducted to examine the performance of these approaches.

Simulation results showed that using parameter constraints allows one to detect various rater effects that are of concern in practice. The Bayesian approach also gave useful results, although estimation of some of the parameters was poor and the standard deviations of the parameter posteriors were large, except when the sample size was large. Using back-reading

scores gave an identified model and simulations showed that the results were generally acceptable, in terms of parameter estimation, except for small sample sizes.

The paper also examines the utility of the approaches as applicable to the PIRLS USA reliability data. The results show some similarities and differences between parameter estimates obtained with posterior mode estimation and with Bayesian estimation. Sensitivity analyses revealed that rater parameter estimates are sensitive to the specification of the priors, as also found in the simulation results with smaller sample sizes.

TABLE OF CONTENTS

Section	Page
Chapter I……..	**1**
INTRODUCTION……	**1**
Challenges of Single Rater Designs ……...	1
Possible Remedies for Identification Issues in Single Rater Designs …….	2
Purpose of Study……..	5
Summary …….	7
Chapter II ……	**9**
LITERATURE REVIEW……...	**9**
1. Approaches to Rater Effects in CR Items …….....	9
1.1. Item Response Theory Models for CR Items ……..…….....	10
Item Response Theory ……..…	10
IRT Models for CR Items …….....	11
The FACETS Model ……	13
1.2 Latent Class Signal Detection Theory ……..…..	14
LC-SDT Model ..…….……	15
Rater Parameters ……...…….....	15
Latent Structures of LC-SDT …….…….	17
Classification Accuracy…….….....	19
2. Issues in Single Rater Designs…….…...……...	19

TABLE OF CONTENTS

2.1. Operational Practice of Single Rater Designs……. 20

2.2. Issues of Model Identification……...…. 22

Identification Issues……...…….. 22

Back-reading and Identification…….. 24

3. Parameter Constraints for Model Identification in Single Rater Designs……. 26

3.1. Equality Restriction…….…….. 26

3.2 Specific Value Restriction…….. 27

4. Bayesian Approaches for Sparse Data ……..…… 29

4.1. Bayes' Theorem and Bayesian Inference …….. 29

4.2. Bayesian Estimation ……... 30

Posterior Mode Estimation (PME) ……. 31

Markov chain Monte Carlo (MCMC) Simulation ……... 31

4.3. Using Priors for Model Identification…….... 33

Informative Priors…… 33

Using Normal Informative Priors ……....……... 34

Priors for PME…….……... 36

Chapter III …… **38**

METHODS …… **38**

1. Simulation Studies ……. 38

Simulation 1: Single Rater Designs …… 38

Simulation 2: Partial Second-Rater Designs …….. 44

TABLE OF CONTENTS

Computational Methods …… 47

Parameter Recovery …… 50

2. Empirical Study…… ……. 51

Chapter IV ……..…… **55**

RESULTS ……..……. **55**

1. Results for Simulation 1 (Single Rater Designs)…… 55

Parameter Estimates ……..……. 56

Standard Errors /Posterior Standard Deviations …… 63

Classification ……. 64

2. Results for Simulation 2 (Partial Second Rater Designs) …….. 67

Parameter Estimates ……..…….. 69

Standard Errors /Posterior Standard Deviations …… 75

Classification ……. 76

Summary of Simulations …….. 78

3. Results for Empirical Study …….. 79

Parameter Estimation ……..….. 80

Sensitivity to Priors ……..…….. 83

Summary of Empirical Study ……..…….. 86

Chapter V ……..……..... **87**

SUMMARY AND DISCUSSION ……..... **87**

TABLE OF CONTENTS

1. Summary and Discussion ……... 87

2. Limitations and Future Research …… 89

REFERENCES……..... **91**

Appendices

Appendix A ……. 100
Outcomes for Parameter Constraints (via PME) in Simulation 1 …… 100

Appendix B …….. 116
Outcomes for Informative Priors (via MCMC) in Simulation 1 ……. 116

Appendix C …….. 132
Outcomes for Bayes' Constants (via PME) in Simulation 2 ……... 132

Appendix D …….. 143
Outcomes for Informative Priors (via MCMC) in Simulation 2 ……. 143

Appendix E …….. 154
Evaluation of the Convergence in MCMC in Simulation (An Example)……. 154

Appendix F …….. 157
Evaluation of the Convergence in MCMC in Empirical Study……...……. 157

LIST OF TABLES

Table	Page
Table III.1……... Data Generation Conditions in Simulations ……	39 39
Table III.2……..... Population Values of LC-SDT Model for the Equal d Condition ……..	41 41
Table III.3……... Population Values of LC-SDT Model for the Varied d Condition …….	41 41
Table III.4……... Simulations using Parameter Constraints…….……	42 42
Table III.5……... An Unbalanced Incomplete Design (10 rater pairs, N=1000, 10% linkage) ……..	46 46
Table III.6……... Average Computational Time for Each Simulation Condition via Openbugs ……	49 49
Table III.7……..... Description of Item 4 and Item 5 in PIRLS 2006 Reliability Data …….	51 51
Table III.8……..... Item 4, Rater Design in PIRLS (*N*=1023, 2nd rater=23%) ……..	53 53
Table III.9……... Item 5, Rater Design in PIRLS (*N*=1000, 2nd rater=20%) …….	54 54
Table IV.1……... Estimated and Obtained Proportion Correct and Correlations with True Latent Classes in a Single Rater Design (Simulation 1) …….	65 65
Table IV.2……. Estimated and Obtained Proportion Correct and Correlations with True Latent Classes in a Partial Second Rater Design (Simulation 2) …….	77 77
Table IV.3……... LC-SDT Parameter Estimates for Item 4 ……	81 81
Table IV.4……... LC-SDT Parameter Estimates for Item 5 ……	81 81

LIST OF FIGURES

Figure	Page
Figure II.1……..……...	16
Distribution for 3 category responses based on LC-SDT……	16
Figure II.2…..…….....….	17
SEM representation of LC-SDT with two raters …….	17
Figure IV.1……...	57
Relative Criteria for PME with Bayes' Constants and Model Constraints (d=3) in Simulation 1 ……	57
Figure IV.2……...	59
Rater Detection Parameters for Bayesian Estimation with Normal Priors in Simulation 1	59
Figure IV.3……...	61
Relative Criteria for Bayesian Estimation with Normal Priors in Simulation 1……..	61
Figure IV.4……...	62
Latent Class Sizes for Bayesian Estimation with Normal Priors in Simulation 1……...	62
Figure IV.5…….....	68
Rater Detection Parameter for PME with Bayes' constants in Simulation 2……..	68
Figure IV.6……...	69
Relative Criteria for PME with Bayes' constants in Simulation 2 …….	69
Figure IV.7……...	70
Latent Class Sizes for PME with Bayes' constants in Simulation 2 ……..	70
Figure IV.8……...	72
Rater Detection Parameters for Bayesian Estimation with Normal Priors in Simulation 2.	72
Figure IV.9……...	73
Relative Criteria for Bayesian Estimation with Normal Priors in Simulation 2 …….	73
Figure IV.10…….....	74
Latent Class Sizes for Bayesian Estimation with Normal Priors in Simulation 2 ……..	74
Figure IV.11…….....	82
Relative Criteria Estimates (c) ……	82
Figure IV.12…….....	84
Effects of Bayes' constants in PME ……...…….	84

LIST OF FIGURES

Figure IV.13……..... 85
Effects of Normal Prior Variances for Bayesian Estimation ……...……... 85

ACKNOWLEDGMENTS

As T. S. ELIOT once said, April IS the cruellest month, which I learned firsthand. Not only did I have a newborn baby, I also successfully defended my dissertation in April 2013. Laying in my hospital bed, I wished I could hop on a time machine if possible. Now, I am so relieved that I got through both the surgery and the defense. I couldn't have done it without help — starting with my advisor Professor Lawrence DeCarlo. He guided me through the dissertation from the beginning and made it possible to finish. I greatly appreciate all of his input and efforts to guide me towards becoming a researcher. His enthusiasm and attitude towards research have always inspired me.

I would also like to thank my other dissertation committee members. Thanks to professor Matthew Johnson, who has been supportive from the beginning and provided brilliant insights into it. Thank you to Professor Young-Sun Lee for serving as the committee chair in such a busy time and for her academic and emotional support throughout this doctoral journey. I also appreciate Professor Victor de la Pena and Professor George Gushue who helped me improve my work.

I am also grateful for my mentors who have helped me become a researcher. Professor Gregory Camilli at Rutgers Univerisity, I really appreciate all of your support, especially in guiding me to the field of measurement and training me from the beginning. Thanks to Professor Myoungsoon Kim in Younsei University, who cared for me since freshman year and helped me enter the research field. I would also like to thank Professor George Bonanno, whom I consulted with for his research, for his warm considerations while I went thru this difficult time. And,

thanks to Professor Bruce Levin for his thoughtful comments during my proposal that helped to improve the entire dissertation.

I also thank my colleagues, Karen, Fran, and Veronica, at the Clinical Education Initiative evaluation center in Columbia University Medical Center, and Jeewha and Nancy at the testing development unit in Oxford University Press for their support and understanding of my requirements to balance school and work. I am very fortunate to have worked with these people while writing my dissertation and learning real world applications. Brian, Chien-mao, Ray, Yoonsoo, Gerald, Zhifen, Jungyeon, and my other classmates, I truly appreciate your help and will miss you all.

I dedicate this dissertation to my family. I am truly grateful to my parents, Junwhan Kim and Yangnim Park, who always believed in me and are the best parents in the world. Thanks to my father- and mother-in-law, Junwon Ko and Soonhee Kang for their support. I also thank my brother's family, Sunghyun and Minhee Kim, and my brother-in-law, Jihoon Ko, for their continuous encouragement. Last but not least, I thank my husband, Jung Hoon Ko, and my two children, Darby and Jason, who were the unfortunate victims but the true winners of my doctoral study. I really appreciate their sacrifices and love.

Chapter I

INTRODUCTION

In many operational assessment situations employing a constructed-response (CR) item (e.g., an essay, an open-ended question, or a performance rating), an examinee's response is evaluated by only one rater. Assigning one rater to each examinee's CR item is called a 'single rater design' (Skyes, Ito, & Wang, 2008). Single rater designs are widely used in low-stakes assessments, such as classroom assessments, as well as some large scale assessments, e.g., TIMSS (Trends in International Mathematics and Science Study) and TOEFL iBT® speaking assessments (ETS). This study examines single rater designs as well as *sparse rater designs*, which include single rater designs as well as designs with some scores from a second rater.

Challenges of Single Rater Designs

In contrast to single rater designs, multiple rater designs are used by many high stakes tests for quality control purposes. For instance, the Advanced Placement Program (AP, College Board) and TOEFL iBT® writing assessment (ETS) uses at least two raters per essay. The rationale for using multiple raters is based on the fact that it is difficult to control the quality of rater scorings in single rater designs. Since raters add a subjective layer to the scoring process, different raters can possibly award different scores to the same response. Therefore, different single raters' subjectivities may lead to systematic scoring biases.

Systematic biases in CR item scoring due to raters, namely *rater effects*, have been widely discussed in earlier literature (e.g., DeCarlo, 2002, 2005; Myford & Wolf, 2003, 2004; Saal, et al., 1980). For example, the average rating for a rater may be lower than the average rating of other raters. This rater 'severity' causes an ambiguity. In particular, in a single rater design, it cannot be determined whether a rater tended to assign systematically lower ratings than

other raters, or whether the set of examinees the rater evaluated tended to be of lower quality in their responses than other examinees.

Few measurement models have been proposed to account for rater effects. However, a latent class extension of signal detection theory (LC-SDT; DeCarlo, 2002, 2005) provides a conceptual framework for the study of rater behavior in CR scoring and is of central interest here. LC-SDT models include parameters regarding raters' performance, such as the raters' ability to discriminate (or detect) latent categories—which indicates the individual rater's reliability — and raters' use of response criteria—which reflects various rater effects that are found in practice.

Similar to other measurement models for rater effects, LC-SDT models require multiple raters per examinee in order to obtain a unique estimate for each model parameter, that is, in order for the model to be identified. For example, for CR items that contain at least three response categories, at least two raters are necessary for the LC-SDT model to be identified (DeCarlo, 2002). In a single rater design, however, only one rater per examinee is available and so LC-SDT models are not identified. This identification issue limits the use of LC-SDT models in a single rater design; for instance, while performances of LC-SDT models in multiple rater designs have been reviewed with CR items in simulation studies (DeCarlo, 2008; 2010), its' utility in a single rater design has never been investigated.

Possible Remedies for Identification Issues in Single Rater Designs

Parameter constraints. One approach to alleviate model identification issues is to simplify the model. This approach has been employed in studies with other models with potential identification problems. For example, in the context of latent class analysis, researchers have suggested constraining model parameters in order to make the model identifiable (de Leeuw, van der Heijden, & Verboon, 1990; Goodman, 1974; Vermont & Magidson, 2005).

For instance, the LC-SDT model parameters (e.g., rater detection) can be restricted to be equal for all of the raters in order for the model to be identified. Another possible restriction would be assigning a specific value to a LC-SDT model parameter, where plausible values for LC-SDT model parameters can be obtained from prior research.

In situations with insufficient observations (e.g., weak identification situations), studies based on item response theory (IRT) have investigated the utility of equality constraints (e.g., Lord, 1983; Parshall, Kromrey, & Chason, 1996), or a restriction that assigns specific values to parameters (e.g., Barnes & Wise, 1991). These studies have found that the simplified models are viable alternatives to the original models.

Bayesian methods. Another approach for addressing identification issues involves Bayesian methods. Several studies have employed Bayesian methods for CR item analyses in various missing data structures (Cap, Stokes, & Zhang, 2010; Patz & Junker, 1999; Lee & Song, 2004; van Onna, 2002). Bayesian inference relies on Bayes' theorem, a mechanism that incorporates *prior* beliefs (or density about parameters and hypotheses learned from the data) to yield *posterior* beliefs or density.

In particular, informative priors have been suggested for models with identification issues by several researchers (Galindo-Garre, et al., 2004; Greenland, Schwartzbaum, & Finkle, 2000; Kass, Carlin, Gelman, & Neil, 1998). An informative prior refers to a prior distribution that gives numerical information that is crucial to parameter estimation and often comes from literature reviews or explicitly from an earlier data analysis.

A commonly used informative prior is the normal distribution. Two parameters of the normal distribution—mean and variance—can be specified to represent a researcher's beliefs or hypotheses about the model parameter. Normal priors are often employed with Markov chain

Monte Carlo (MCMC) estimation methods (due to the complexity of the posterior density). The application of informative priors with normal priors is somewhat similar to assigning a specific value restriction on the model parameters, except that Bayesian methods allow one to have uncertainty about possible values. Therefore, it is worthwhile to explore the utility of informative normal priors via MCMC for LC-SDT model parameter estimation in sparse rater designs.

In addition, the use of Bayes' constants and posterior mode estimation (PME) has been suggested by several authors (Galindo-Garre & Vermunt, 2006; Schafer, 1997; Vermont & Magidson, 2005) to deal with situations with estimation problems. In PME, only the mode of the posterior is obtained instead of the full posterior distribution; Bayes constants are hyperparameters of Dirichlet priors that are used for the conditional response probabilities and the latent class probabilities (see DeCarlo et al., 2011). In cases with model identification problems, prior information via Bayes' constants could provide just enough information to uniquely determine the parameter values. The estimation method of Bayes' constants via PME has been primarily employed in previous studies on LC-SDT models (without identification issues) and by Galindo-Garre and Vermunt (2006), who reported benefits of this method over a fully Bayesian analysis. Hence, the current study also examines the utility of PME in sparse rater designs.

Back readings. To remedy identification issues, one can also use second rater back-readings, which are occasionally collected for a subset of examinees in single rater designs. For example, in a study with the AP® English literature and composition examination (Wolfe, Myford, & Englehard, Jr., 2007), rater group-leaders reviewed selected essay ratings, which is 'back reading'.

Having partial second rater scores creates a data structure where most of the examinees only have one rater score and the majority of the second rater's scores are missing by design. For example, the scoring rules for NETP (National Education Technology Plan, 2010) include 10% back-readings for simple mathematics items (Jones & Vickers, 2011), and so 90% of the second rater's scores are missing. This type of missingness is ignorable (Rubin, 1976), so it does not cause biased estimation.

However, this situation often leads to large standard errors for the parameter estimates, due to 'weak identification' (Vermunt & Magidson, 2005). To counteract the low number of raters per examinee, a large dataset (i.e. a large sample size or a higher proportion of second raters), is needed in order to obtain adequate estimates of model parameters (DeCarlo, 2002; DeCarlo & Kim, 2008).

Purpose of Study

The purpose of the proposed study is to investigate possible solutions for LC-SDT models in sparse rater designs. Particularly, the present paper explores the utility of parameter constraints and Bayesian methods in situations with model identification issues. The use of back-readings is also examined. The performances of the proposed approaches are compared with respect to parameter recovery and classification accuracy in both simulations and empirical studies.

Simulation study. Simulation studies are conducted to examine the utility of parameter constraints, Bayesian methods, and back-readings in sparse rater designs. Simulations are designed to investigate these approaches in different conditions by manipulating sample size, rater effects (e.g., severity-leniency), and rater discrimination. The simulation studies examine the following:

(1) the ability of LC- SDT models to detect rater effects (e.g., shifted rater criteria) in a sparse rater design,

(2) the impact of sample size on parameter recovery and classification,

(3) the impact of parameter constraints, Bayesian approaches, and back-readings on parameter recovery and classification, and

(4) the effect of Bayesian informative priors in sparse rater designs.

Because rater effects are a source of systematic error in performance ratings, it is important to monitor rater effects and to adjust for these effects. An advantage of the LC-SDT model is that it allows one to do just that. Hence, the performance of LC-SDT models in sparse rater designs with rater effects is examined. Specifically, simulations are used to examine the ability of the model to detect shifted rater criteria (e.g., rater effects) in sparse rater designs. For example, if the criteria are all shifted upwards, then raters are stricter, because they tend to give lower scores. If they are shifted downwards, then raters are more lenient. Bayesian estimation methods (MCMC with normal priors) and PME with Bayes' constants are examined.

The simulations also include situations where the true values of rater discrimination parameters are misspecified in the fitted model, in order to determine the effect on recovery, rather than simply examining the situation where the true values and researcher-inputted detection values are the same. The simulations also investigate the impact of sample size on parameter recovery. In addition, the impact of the percentage of partial second rater scores that are available is examined to provide some guidelines for practice.

Empirical study. In the empirical study, LC-SDT models are applied to real-world data, specifically PIRLS (progress in international reading literacy study; Mullis, Martin, Kennedy, & Fox, 2006). The PIRLS data that are examined are reliability datasets which have partially

available second rater back-readings (which is why they were used). The analysis consists of the following:

(1) comparing results across different estimation methods,

(2) seeing if different rater effects can be detected, and

(3) examining the effects of priors on model estimation.

The utility of two types of estimation methods— Bayesian estimation with normal priors and PME with Bayes' constants—are examined in terms of parameter recovery, standard deviations, and computational time.

Patterns of rater effects are summarized in terms of the latent class SDT model. Rater effects are examined using plots of parameter estimates reflecting rater criteria and rater discrimination. In particular, rater criteria parameters are investigated for rater severity-leniency and rater discrimination parameters are examined to assess rater accuracy.

Since Bayesian methods are based on the specification of priors, it is beneficial to examine how these priors impact parameter recovery, especially in sparse rater designs where the data are sparse or only small sample sizes are available. The investigation examines effects of using different Bayes' constants in PME and of using different normal priors in MCMC. A sensitivity analysis is conducted to examine the effects of these priors parameter estimation.

Summary

The paper explores possible approaches for using LC-SDT models in sparse rater designs by examining three approaches: 1) the use of parameter constraints, 2) the use of Bayesian methods with informative priors, and 3) the use of back-readings. In order to investigate these approaches, simulations are used to examine parameter recovery and classification accuracy. The

approaches are also applied to real-world data: the PIRLS reliability datasets which have partially available second rater back-readings.

Chapter II discusses approaches to various rater effects including item response theory and of course the LC-SDT model. Then, previous research regarding sparse rater designs is reviewed along with approaches that deal with model identification issues and insufficient observations. In Chapter III, methods for assessing the proposed approaches are outlined for the simulations and empirical studies. Chapter IV presents results of the simulation studies as well as real world data analysis. Finally, Chapter V summarizes findings of the study and discusses implications of the results and limitations of the study.

Chapter II

LITERATURE REVIEW

This chapter begins with a review of issues in sparse rater designs. First, Section II.1 presents approaches to rater effects via various measurement models including the LC-SDT model. Section II.2 describes potential issues in modeling rater effects in single rater designs, then possible solutions that have been applied in similar situations are discussed. In particular, Section II.3 reviews model modification methods, and then Section II.4 introduces the fundamental idea of Bayesian methods which are explained in detail.

II.1. Approaches to Rater Effects for CR Items

Given recent attention to constructed-response (CR) items for psychological and educational measurement, there has been an increased demand for scientific tools to investigate the performance of raters. For instance, with consideration for the complex nature of performance ratings involving human judgment, *the Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) requires performance assessment developers to monitor and report scoring errors and to correct any systematic sources relating to these errors.

One of the most common methods used to review rater performance in medical and social science research is the analysis of inter-rater agreement (or disagreement). Numerous extensions and generalizations of this inter-rater agreement measure have been proposed in the literature (e.g., Agresti, 1992; Cohen, 1960; Tanner & Young, 1985; Uebersax & Grove, 1990). However, limitations of rater agreement statistics have been discussed in previous literature (Agresti, 2002; Banerjee, et al., 1999; DeCarlo, 2002; Uebersax, 2012b) and the artificiality of emphasizing rater

agreement has been questioned by rater training practitioners (Elder, Barkhuizen, Knoch, & von Randow, 2007; Weigle, 1998).

Moreover, many studies have found that the performance of raters in CR item scoring differ considerably across raters (Engelhard, 1994; Englehard & Myford, 2003). Without any adjustment for this rater variation, simple raw ratings might give invalid estimates for an examinee's performance qualities. Given these criticisms, various modeling approaches have been marshaled to monitor rater effects and to adjust for these effects on scoring (see, e.g., Agresti, 2002; DeCarlo, 2002; von Eye & Mun, 2005; Tanner & Young, 1985).

In the following sections, previous approaches to monitor rater effects on CR items are reviewed, starting with a popular measurement model that is widely used to analyze CR items and then moving on to the LC-SDT model.

II.1.1 Item Response Theory Models for CR Items

Item response theory (IRT) has been employed in measurement for the analysis of CR items. The FACETS model (which is a commonly employed in rater effect investigations) is a type of IRT model. Hence, this section reviews the basic concepts of IRT, relevant IRT models that account for CR items, and the FACETS model.

Item Response Theory

Item response theory (IRT) provides a framework for estimating an examinee's latent proficiency or ability to account for item effects and for monitoring item performance (Embretson & Reise, 2000). In typical IRT, a model is formulated to explain the probability of a correct answer. Let Y_{il} denote the response of person i on item l (l=1,…,L). In case the items are dichotomously scored as incorrect or correct, Y_{il} = 0, 1 respectively, then, the probability of a

correct answer is defined by the parameters of interest including the person's latent ability θ_i and item parameters (item discrimination a_l and item difficulty b_l) such that,

$$p(Y_{il}=1|\theta_i) = F[a_l(\theta_i - b_l)]. \quad \text{(II.1.1)}$$

Here, a latent score θ_i can be used to estimate an examinee's ability, and it is often assumed to be distributed as standard normal with mean of 0 and *SD* of 1 for model identification. Item difficulty b_l represents the point on the ability scale where an examinee has a 50% chance of answering item *l* correctly, thus, a high b_l indicates a hard item. Item discrimination a_l determines the rate of change in the probability of answering an item correctly as a function of θ_i, and items with higher discriminations are more useful for separating examinees into different ability levels.

The most commonly used cumulative density function (CDF), e.g., *F* in Equation II.1.1, in IRT are logistic or normal density functions. With a logistic CDF, the model in Equation II.1.1 is referred to as the two-parameter logistic (2-PL) model. When the item discrimination parameter a_l is set to be equal across items, the 2 PL model simplifies to a one parameter (1-PL) model, or equivalently the Rasch model (Rasch, 1960).

IRT Models for CR Items

For CR items where the observed responses are on an ordinal scale, IRT models for polytomous responses have been used. Among them, the graded response model [GRM] (Samejima, 1969) uses cumulative logits for each response category. For ordered response category *m*, the GRM can be written as,

$$\log[\frac{p(y_{il} \le m \mid \theta_i)}{p(y_{il} > m \mid \theta_i)}] = a_l(\theta_i - b_{lm}). \quad \text{(II.1.2)}$$

Note that the item difficulties, b_{lm}, are defined to differ across response categories and are referred to as "category threshold parameters", whereas the item discrimination parameter is

assumed to be invariant across different response categories for a specific item *l*. This model is similar to an LC-SDT model, where item discrimination and category threshold parameters are analogous to rater detection and rater criteria parameters, respectively. However, latent ability, θ_i, in the GRM is continuous, whereas the LC-SDT model contains a discrete latent variable. Hence, the LC-SDT model can be viewed as a "semi-parametric version of GRM" (DeCarlo, 2005, p.59).

Another commonly used polytomous IRT model is the generalized partial credit model (GPCM) by Muraki (1992) based on the partial credit model (PCM) proposed by Masters (1982). The GPCM can be written as,

$$\log[\frac{p(y_{il}=m\,|\,\theta_i)}{p(y_{il}=m-1\,|\,\theta_i)}]=a_l(\theta_i-b_{lm}), \qquad \text{(II.1.3)}$$

where b_{lm} is now an item step parameter, which indicates the item difficulty for moving from category m–1 to category m. The PCM is a special case of the GPCM, where item discrimination a_l is assumed to be equal for all items. GPCM utilizes adjacent-category logits, that is, it compares $p(y_{il}=\mathrm{m}|\theta_i)$ to $p(y_{il}=\mathrm{m}-1|\theta_i)$, unlike the GRM with cumulative logits, which compare $p(y_{il}\leq \mathrm{m}|\theta_i)$ to $p(y_{il}>\mathrm{m}|\theta_i)$. Also, it is allowed in the GPCM that item step parameters are not ordered, while GRM assumes strictly ordered item category threshold parameters.

These IRT models provide better estimates for an individual's ability by accounting for item bias, rather than simply aggregating item-total scores. However, the utility of these models are limited for CR items that require scoring by raters. For example, it has been noted that item parameters within these IRT models are confounded with rater effects (Boughton et al., 2001; DeCarlo et al., 2011). Regarding this issue, an extended IRT approach, the FACETS model, (Linacre, 1996) has been widely employed for investigating rater effects in previous literature (e.g., Engelhard,1994; Myford & Wolfe, 2003, 2004).

The FACETS Model

The FACETS model is an extension of the Rasch model and includes additive effects, e.g., rater FACETS, in addition to item and examinee FACETS on a logit scale. The FACETS model for CR items can be written as,

$$\log[\frac{p(y_{il}=m\,|\,\theta_i)}{p(y_{il}=m-1\,|\,\theta_i)}]=\theta_i-b_l-\gamma_m-\xi_j, \tag{II.1.4}$$

where b_l indicates the item difficulty, and γ_m is the item step parameter, thus, the first three terms in Equation II.1.4 are equivalent to a PCM in Equation II.1.3. The additional parameter ξ_j represents rater severity, which indicates the tendency of raters to be lenient or strict.

Several studies (e.g., Saal, et al., 1980; Myford & Wolfe, 2003, 2004) portray rater effects as the source of systematic error in performance ratings via the FACETS models. They focus on particular patterns relevant to a rater's use of criteria. Some of those rating patterns are labeled as "severity" or "leniency" (i.e., a rater rates above or below the average rate), centrality or extremism (i.e., a rater overuses or avoids the extreme categories), or range restriction (i.e., a rater overuses any point on a rating continuum).

However, Saal, et al.(1980) and Myford and Wolfe(2003) discussed limitations of prior research on rater effects via FACETS models, such as the fact that a separate Facet model is required to monitor a particular rater effect. For example, a FACETS model with a severity parameter can only show rater severity or leniency (via a rater severity parameter representing the average severity across response categories for a rater) and centrality, but not any other rater effects.

Moreover, the FACETS model assumes all raters have equal detection or discrimination. This assumption can lead to estimates of the item difficulty parameters that are biased. For

instance, Patz, et al. (2002) found that with the FACETS approach, the item difficulty parameter estimates are shrunk towards zero.

Additionally, several researchers (Donoghue & Hombo, 2000; DeCarlo et al., 2011; Mariano, 2002; Verhelst & Verstralen, 2001) have pointed out that the FACETS model's independence assumption for multiple ratings within an examinee-item pairing is incorrect. Further, DeCarlo et al. (2011) and others (Patz et al. 2002; Mariano, 2002) pointed out that the FACETS model also implies that more precise measurement of an examinee's proficiency can be obtained simply by increasing the number of raters per item, regardless of the number of items—which is rather unrealistic. In particular, Mariano (2002) mathematically showed that in FACETS models, the measurement errors for the examinee's latent proficiency go to zero as the number of raters increases.

II.1.2 Latent Class Signal Detection Theory (LC-SDT)

Latent structure approaches, e.g., latent trait or latent class modeling, provide information about the accuracy of ratings when a "gold standard" is not available (Hui & Zhou, 1998; Walter & Irwig, 1988; Uebersax & Grove, 1990). Specifically, latent class analysis (LCA) is a statistical technique for characterizing latent classes for categorical or ordinal data (e.g., CR items). LC models (e.g., Agresti, 2002; Uebersax & Grove, 1990) treat both the observed scale and the latent variable as categorical or ordinal, which is more appropriate for CR items, because they use (latent) categories in the scoring rubrics (e.g., novice, intermediate, or advanced).

However, general latent class models that account for rater agreement (e.g., Murphy & Balzer 1989; Qu, Tan, & Kutner, 1996; Uebersax, 1999) have been criticized with regards to the meaning of model parameters (Nelson & Pepe, 2000); hence, DeCarlo (2002) introduced a latent class version of signal detection theory, which provides a well-developed psychological

interpretation, based on signal detection theory (SDT), about what raters do when scoring CR items.

LC-SDT Model

The LC-SDT model can be written as,

$$p(Y_{ij} \leq \mathrm{k} \mid \eta_i) = F(c_{jk} - d_j\eta_i), \qquad \text{(II.1.5)}$$

where Y_{ij} is the response variable for a rater j (j=1,…J) assigning discrete score k (k=1, …, K response categories) to examinee i's (i=1, …, N) response on a CR item. The parameters of interest are 1) c of rater criteria, 2) d of rater detection and 3) η as the 'latent' or 'true' category of the CR item (η=0,…, K–1). For a cumulative logistic distribution of F, c and d indicate the intercepts and the slope for a latent categorical variable η in an ordinal logistic regression model. The following sections discuss the rater parameters and the latent structure of the LC-SDT in detail.

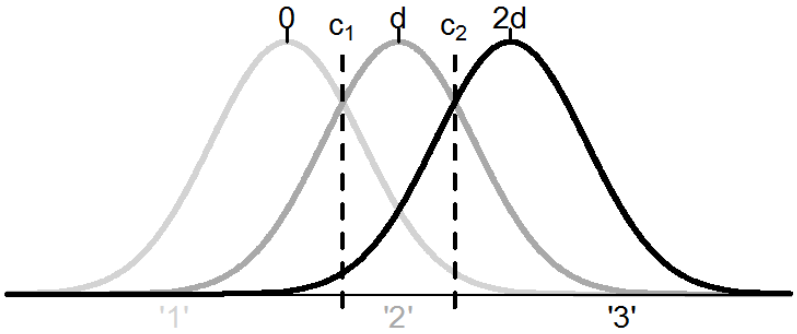
Rater Parameters

Rater detection. The LC-SDT approach to CR item scoring summarizes two basic aspects of the raters' behavior (DeCarlo, 2008, p.2): a *perceptual* aspect and a *decision* aspect. The perceptual aspect refers to the view that the raters' ratings are in part based on their *perception* on the overall quality of an essay (for holistic scoring). This perceptual aspect is reflected by the detection parameter d, which indicates the rater's ability to discriminate between the latent categories.

In Equation II.1.5, with an assumption of proportional odds, the cumulative odds ratio (e.g., the detection parameter) is set to be equal across the latent categories. This equal spacing SDT model (DeCarlo, 2002, 2005, 2008) assumes that a rater perceives the latent classes as being equally spaced, and so, the distance between perceptual distributions is the same for

adjacent distributions, which gives distances of *d*, 2*d*, and so on, as shown in Figure II.1. This restriction is implemented in the model by scoring the latent classes as 0, 1,…,*K*–1 for response categories 1, 2, … , *K*.

Figure II.1. Distributions for 3 category responses for LC-SDT



When the rater has good discrimination between the latent classes (indicated by a large the distance between above distributions), the rater's perceptions of each scoring category are well separated. This means that there is small overlap between the distributions and small error in terms of a rater's ability to classify the CR item.

Rater criteria. A rater's use of criteria for each latent category reflects the *decision* aspect of the task and is illustrated by criteria parameters *c* in LC-SDT. For example, a rater's decision of Category 1 as opposed to Category 2 is reflected by c_1 in Figure II.1. If the rater's perception of a CR item falls below the criteria c_1, a rating of 1 would be given.

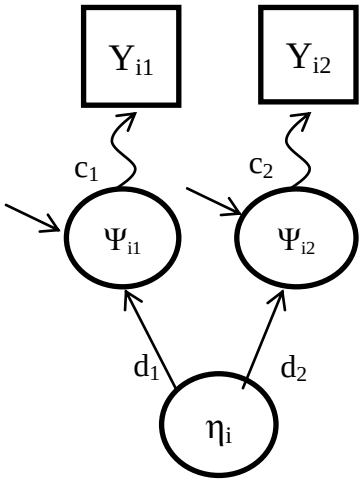
The locations of the response criteria c_k reflect the rater's category usage. For example, previously described rater effects for the FACETS models simply reflect the raters' arbitrary use of response criteria, which are lower (i.e., further to the left) for lenient raters and higher for strict raters, in terms of SDT. By allowing for differences in *c*'s among raters, the locations of *c*'s in LC-SDT also can capture other rater effects, such as centrality or extremism (i.e., a rater avoids or overuses the extreme categories), or range restriction (i.e., a rater does not use all the categories of the scale).

The advantage of LC-SDT models over FACETS models lies in the fact that rater effects beyond simply severity or leniency can be accounted for in LC-SDT, which is not the case for FACETS models. In several real-world applications, LC-SDT models have revealed various rater effects (e.g., DeCarlo, 2008; DeCarlo, 2010; DeCarlo et al., 2011).

Latent Structure of LC-SDT

The structure of a LC-SDT model can be represented as a type of structural equation model, as was noted by DeCarlo (2008). With observed CR scores (e.g., Y_{ij}), the models can be motivated by assuming that there is a continuous underlying variable Ψ_{ij}, which is a rater's perception of the overall quality of an examinee's constructed response, such as essay. Figure II.2 shows a representation of LC-SDT for the case of two raters scoring a single CR item.

Figure II.2. SEM representation of LC-SDT with two raters and one CR item.



In Figure II.2, a rater's use of criteria is represented by c_j. The curved arrows reflect that there is a nonlinear relationship between Y_{ij} and Ψ_j, given that the model applies to the probability of Y_{ij} and not the magnitude of Y_{ij} (i.e., the model links the mean of Ψ_{ij} to the response probabilities). For an example with a dichotomous response, the probability of observing the lowest level category (e.g., k=1) given the lowest level latent category (e.g., η_i=0) can be estimated by the area under the density curve of Ψ_j that is below c_j,

$$p(Y_{ij}<1|\,\eta_i=0) = p(\Psi_{ij} < c_j|\,\eta_i=0) = F(c_j). \quad \text{(II.1.6)}$$

Also, a rater's perception of the different latent classes is given by a structural component in LC-SDT, such that,

$$\Psi_{ij} = d_j\eta_i + \varepsilon_{ij}, \quad \text{(II.1.7)}$$

where ε_{ij} denotes (perceptual) error. It should be noted that a linear relationship of Ψ_{ij} with η_i is assumed and is shown as straight lines in Figure II.2. This means that the mean of the Ψ_{ij} distribution is shifted by d_j across the latent classes, η_i. By combining Equation II.1.6 and Equation II.1.7, it can be shown that the general LC-SDT model specification of Equation II.1.5 follows.

Along with the SDT parameters, e.g., *c* and *d*, LC-SDT models also include a higher level component. The model is basically a restricted latent class model for the joint probability of the response patterns ($k_1, k_2,\ldots, k_J$), which can be written as,

$$p(Y_1=k_1,\ldots, Y_J=k_J) = \Sigma_\eta\, p(\eta)\, p(Y_1=k_1,\ldots, Y_J=k_J|\,\eta), \quad \text{(II.1.8)}$$

where the summation is over the latent classes η, $p(\eta)$ is the probability (size) of the latent class and $p(Y_1=k_1,\ldots, Y_J=k_J|\,\eta)$ is the conditional probability for the response patterns conditional on η. With an assumption of local independence, the second term on the right becomes

$$p(Y_1=k_1,\ldots, Y_J=k_J|\,\eta) = \Pi_j\, p(Y_j=k_j|\,\eta), \quad \text{(II.1.9)}$$

where the product is over the *J* raters. To obtain $p(Y_j=k_j|\,\eta)$, the cumulative probabilities can be differenced, such that,

$$p(Y_j=k\,|\,\eta) = p(Y_j\leq k\,|\,\eta) - p(Y_j\leq k-1|\,\eta), \quad \text{(II.1.10)}$$

where $p(Y_j\leq 1\,|\,\eta) = 0$ and $p(Y_j\leq K\,|\,\eta) = 1$, and then this probability is given by the LC-SDT model in Equation II.1.5.

Classification Accuracy

Several classification indices are also available, as discussed in the literature on latent class analysis (e.g., see Clogg, 1995; Dayton, 1998). In latent class analysis, the cases are classified by using the posterior probability of η given the observed response pattern, which can be written as $p(\eta| Y_1=k_1,\ldots, Y_J=k_J)$. Given estimates of the LC-SDT model parameters, such as $p(\eta)$, c, and d, one can classify the cases and obtain an estimate of the expected proportion of cases correctly classified (Clogg, 1995), P_C. P_C is basically a weighted mean of the maximum posterior probabilities for each unique response pattern and indicates the expected proportion of cases that are correctly classified by the model. One minus P_C gives the classification error rate.

Another classification index is λ (Goodman & Kruskall, 1954), which provides a correction for "chance" to P_C; it gives the relative reduction in classification errors. Values of lambda greater than zero indicate that using the posterior probabilities to classify cases gives an increase in classification accuracy over and above that obtained by simply assigning all cases to the latent class with the largest size.

LC-SDT models have also been employed and expanded upon in several studies concerned with CR scoring (e.g., a hierarchical LC-SDT model for rating multiple items; DeCarlo, 2010; DeCarlo et al., 2011). However, it's utility for the situation where only one rater is assigned to score a CR item has never been investigated. The next chapter introduces the challenges of applying LC-SDT models in that situation, namely sparse rater designs.

II.2. Issues in Single Rater Designs

The study of rater behaviors in CR item scoring is often complicated by the particular methods used to assign raters to constructed responses (e.g., Hombo, Donoghue, & Thayer, 2000; Park, 2011; Skyes, et al., 2008). While a fully crossed design (i.e., all raters score all essays) is

the ideal approach, many large-scale assessments employ only one rater for each examinee's response because of cost and time limitations.

II.2.1. Operational Practice of Single Rater Designs

Single rater designs (Skyes, et al., 2008), or nested designs (Brennan, 1992; Hombo et al., 2001), are one of the most commonly applied methods in assigning raters to operational assessment situations. It refers to the process where one rater (rather than multiple raters) is assigned to evaluate each examinee's response.

For example, the operational condition for the writing and speaking modules in IELTS is a 'one-to-one format' (Taylor & Jones, 2001), i.e. one examinee and one examiner interact throughout all the items within the module. Similarly, the CR items in TIMSS and PIRLS are graded by one rater per examinee within the entire test-set. While the TOEFL iBT® speaking section assigns different raters for different items within the section, and so an examinee's response on an item is scored by only a single rater.

One reason for the use of single rater designs is the resources spent on human raters. Since scoring CR items requires raters, multiple raters for each CR response will lead to time and expense increases involved in scoring, such as rater selection, rater training and qualification, and rater monitoring. Considering these substantial costs for raters, single rater designs are the most cost-effective and efficient method among all rater designs (Sykes et al., 2008).

One justification for single rater designs is an adequate level of reliability. In the IELTS speaking module, Taylor and Jones (2001) found that the reliability of the test via generalizability theory (g-theory) is within a reasonable range (e.g., g coefficient of 0.86 with 4 items). However, this reliability for a single rater design was lower than that for 9 raters, and the

reliability coefficient decreased as the number of items was reduced from 4 to 1 (g coefficient of 0.75).

Another study regarding the IELTS writing module (Shaw, 2004) reported similar levels of g-coefficients over four training sessions for raters, but the multiple rater-agreement level (Kappa, which adjusts for chance agreement) for the last training session was 0.45. This level of agreement can be considered as moderate to poor, but the author mentioned that the rater-agreement level was even lower for earlier training sessions (that showed reasonable g-coefficient levels). These results demonstrate the limitation of reliability in single rater designs; e.g., g-coefficients only indicate that raters use the scoring rubric in generally the same way.

Moreover, since raters introduce a subjective element into the scoring process, different raters can award different scores to the same response. Thus, single rater designs raise issues with respect to the subjectivity of single raters. For example, the previously mentioned study on the IELTS writing section (Shaw, 2004) reported evidence of differences in rater severity (i.e., the average rating for a rater being higher or lower compared to other raters).

A consequence of ignoring this variation in rater behaviors was demonstrated in a simulation study (Hombo et al., 2001). By applying IRT models (and ignoring the rater effects) in a single rater design, the authors found that the errors for examinee ability estimates were large, especially for examinees with extreme ability levels.

A few measurement models have been proposed to take into account different raters' behaviors, e.g., FACETS and LC-SDT models. However, these models are not applicable in single rater designs because of model identification issues. A review of model identification issues is provided in the next section.

II.2.2. Model Identification

Identification Issues

Identification is relevant to all measurement models. In applying these models, a researcher needs to establish whether unique estimates exist for all of the parameters in the model in order for the model to be identified. Bollen (1989) suggested that identification can be indicated by the fact that "the unknown parameters are functions only of the identified parameters *and* that these functions lead to unique solutions (p. 88)." Situations of non-identification occur when there are fewer equations than unknowns. A simple example can be found in Kenny, Kashy, & Bolger (1998).

> For a model in which two variables indicate a single latent variable, there is a single correlation. With that one correlation, it is impossible to solve for the two factor loadings. The problem is a standard algebraic one of fewer equations than unknowns. In cases such as this, there is no unique mathematical solution for the model parameters and the model is said to be not identified. (p. 253)

This example demonstrates the fundamental concept of identification: for a given value of one equation (e.g., the correlation), there exists an infinite set of two unknown parameters (of factor loadings).

Non-identification often implies that different parameter estimates yield the same log-likelihood value in maximum likelihood estimation (MLE). The likelihood function contains the unknown parameters from the model and the observations from the sample; hence, in the application of MLE, an identified model requires more observations (equations) than model parameters (unknowns).

Model identification can sometimes be determined by a formal mathematical analysis. For instance, Bollen(1989) provides several rules of thumb that are *sufficient* or *necessary* for

model identification (e.g., *t*-rule: the number of non-redundant elements in the covariance matrix of the observed variables must be greater than or equal to the number of unknown parameters). While his methods are useful for some typical measurement models (e.g., factor analysis), it is difficult to apply those rules in models with complex structures, such as latent class models.

An empirical identification check for models can be conducted by using the Fisher information (or the information) matrix (Vermunt & Magidson, 2005). The information function indicates the negative of the expectation of the second derivative of the log-likelihood function with respect to the model parameters. When a model is identified, the information matrix should be positive definite. Some of the properties of being positive definite can be used to determine if the model is identified (e.g., all eigenvalues are positive or the information matrix is full rank).

A similar method can be used based on the Jacobian matrix of partial derivatives (Goodman, 1974; McHugh, 1958; Vermunt & Magidson, 2007). A Jacobian matrix consists of all first partial derivatives of the likelihood function with respect to the model parameters. If the rank of the matrix is equal to the number of model parameters, then parameters in the model are 'locally identifiable' (Goodman, 1974). Local identification indicates that conditional on a particular set of parameters, a model parameter is unique in a small neighborhood in the parameter space. While local identification does not prove global identification, it is a necessary condition for global identification (Vermunt & Magidson, 2007).

In practice, statistical software is commonly used to determine whether or not a given model is identified (Hayduk, 1987). Vermunt and Magidson (2005) also suggested that one can check identification by using several sets of starting values; an identified model provides the same unique estimates for different sets of starting values.

While there are no general rules regarding the identification of models with latent classes, one can pin-point certain minimal requirements for model identification (Vermunt & Magidson, 2005). For instance, DeCarlo (2002) noted that, for the LC-SDT model, if raters use at least three response categories, then the model is identified if at least two raters are used. In a single rater design, only one rater per examinee is available, hence, LC-SDT models are not identified.

Back-reading and Identification

Back-readings from second raters can alleviate the model identification issue since those observations can be used to satisfy the minimum requirement of two raters in LC-SDT models with CR items. In many assessment practices involving single rater designs, a second rater is assigned for partially selected constructed responses to obtain 'back-reading'; this is usually done in order to estimate rater-agreement statistics.

For example, in the scoring procedure of the IELTS writing and speaking modules, selected centers provide a representative sample of examiners' responses in marked tapes and scripts, which are then back-read by a team of IELTS Principal Examiners and Assistant Principal Examiners. The data is used for the analysis of the paired examiner–Principal Examiner ratings to examine the quality of the ratings. Similarly, in TOEFL iBT® speaking, partial double scoring is conducted at each test administration for quality control and rater monitoring.

However, these partial second rater observations create a sparse data structure (i.e., many zero frequencies for possible observations), where most of the examinees only have one rater score and the majority of the second rater's ratings are missing by design. This type of missingness is ignorable (Rubin, 1976), and so it doesn't result in biased estimation. However, the situation of minimal possible raters with high missing rates often leads to 'weak identification' issues (Vermunt & Magidson, 2005).

Weak identification as well as empirical non-identifiability (Uebersax, 2012a) indicates situations where the observations are not informative enough to obtain stable parameter estimates even if the parameters are uniquely determined. It can be detected by the occurrence of large standard errors (Vermunt & Magidson, 2005) or when it takes an unusually long time for convergence to occur (Uebersax, 2012a). Several studies have also investigated consequences of sparseness in categorical data analysis. For example, Agresti and Yang (1986) and Agresti (2002) reported that for sparse data, MLE produced severe biases (e.g., infinite estimates for logit model coefficients). The authors also pointed out that chi-squared approximations for goodness-of-fit statistics were poor for sparse data. Sparseness of the data produces response patterns with zero or very low frequencies, which affects appropriate reference distributions (e.g., chi-squared distribution). Other research has reported low power for the odds ratio test in sparse data situations (Fleiss, Levin, & Paik, 2003).

To counteract identification issues due to few raters per examinee, one can employ a large sample size. For example, DeCarlo and Kim (2008) showed, in a simulation study, that rater parameter recovery for a third rater with 92% missing data was excellent for a sample size of 20,000. This type of sample size is obtained in some large scale assessments. The study also reported benefits for classification accuracy of using sparse third rater observations, as compared to classification accuracy obtained without the third rater observations.

Other possible approaches to model identification include using parameter constraints or incorporating previous knowledge via Bayesian approaches. The next section reviews those approaches.

II.3. Parameter Constraints for Model Identification in Single Rater Designs

This section reviews literature regarding other approaches to model modification that can be used for single rater designs. In particular, two types of restrictions for simplifying the model are discussed: equality restrictions and specific value restrictions.

II.3.1. Equality Restrictions

Equality restrictions assume that some parameters are equal, which simplifies the model by reducing the number of unknown parameters and leads to an identified model. For example, in the previous example with two indicators for a one factor model (Kenny, Kashy, & Bolger, 1998), an equality restriction (e.g., setting two factor loadings equal) adds a second equation and would ensure model identification. This model is often called the 'tau-equivalent' model (Bollen, 1989).

For an unrestricted latent class analysis, Vermunt and Magidson (2005) showed that at least three indicators are needed for identification, but if the indicators are only dichotomous, then no more than two latent classes can be identified. However, the authors illustrated that with the equality restrictions of Pr(response = 1|latent class = 1) = Pr(response = 2|latent class = 2), a two-class model with two dichotomous indicators can be identified. Similar examples of equality restrictions can be found in other studies that employ latent class models (e.g., de Leeuw, van der Heijden, & Verboon,1990; Goodman, 1974; McHugh, 1958).

Another example of the use of equality restrictions can be found in item response theory (IRT). Given a small sample size, an IRT model with two item parameters (e.g., item difficulty and discrimination) would be subject to empirical identifiability issues. In that case, one can use an IRT model with only one item parameter (e.g., item difficulty) and assume that the item discrimination parameters are equal, which gives the Rasch model (Rasch, 1960). Similarly, one can assign an equality restriction on the item guessing parameters in an IRT model with three

item parameters. The performance of these modified models was investigated by Lord (1983) for an equal restriction on item discriminations and by Parshall, Kromrey, and Chason (1996) for an equal restriction on item guessing parameters.

II.3.2. Specific Value Restrictions

Another approach to model identification is to assign specific values to some model parameters. This approach is commonly used, for example, in multinomial regression models, in which the regression coefficient for the last (or the first) category is set to '0' for model identification.

In the example of two indicators for one factor (Kenny, Kashy, & Bolger, 1998), the number of unknowns can be reduced by specifying a value (e.g., 1) for the first factor loading, which may lead to model identification. However, Bollen (1989) showed that even with this specific value restriction, the model is still not identified. He suggested that if we know the reliability of one indicator, with a specific value for the first loading (e.g., 1) the model is then identified. In particular, the variance of the latent variable (e.g., Φ) can be estimated by the product of the reliability and the variance of the indicator (e.g., $\rho \times \text{Var}(x_1)$), and the factor loading for the other indicator can be estimated by dividing the covariance of two indicators by the variance of the latent variable (e.g., $\text{Cov}(x_1, x_2)/\Phi$). The approach used here of setting values for the discrimination parameter (which is related to reliability) is similar to Bollen's suggested approach.

In the application of IRT models to small sample sizes, researchers have employed specific value restrictions, especially for the item guessing parameter. Given IRT models with three item parameters, Barnes and Wise (1991) found problems with convergence in simulations with small sample sizes (50, 100, and 200) and moderate test lengths (25 and 50 items). The

authors found that if they fixed the item guessing parameter (e.g., to 0.2 or 0.25), model convergence improved and ability estimates were more highly correlated to their true values. Setiadi (1997) also found similar results in simulations with larger sample sizes (100, 200, and 500) for two sampling distributions of ability (e.g., normal and uniform).

Similar to the above approaches, the LC-SDT model can be made to be identifiable by assuming that the rater detection parameters are equal for all of the raters or by assigning specific values to the model parameters. For example, it can be assumed that all of the raters have an equal detection parameter, which is similar to assuming that the rater reliability is known. This modification allows one to estimate the rater criteria parameters and therefore still provides information about rater effects. Another possible modification is to fix the latent class sizes, which is similar to assuming a known distribution (i.e., the standard normal) for the ability distribution in IRT. This approach is useful when there is information about likely latent class sizes from the previous studies.

A combination of these two approaches can also be used. That is, rater detection parameters as well as latent class sizes can be treated as known; hence, only the rater criteria parameters are estimated. This case is similar to the 1PL IRT model, which assumes that all the item discriminations are equal and the latent proficiency distribution is standard normal, and only the item difficulty levels are estimated. The number of parameters for a LC-SDT model with these modifications are reduced to $J\times(K-1)$, for rater criteria parameters only, which is equal to the number of the unique observations, and so this approach can be used for single rater designs. The utility of this approach for uncovering rater effects will be examined here.

II.4. Bayesian Approaches for Sparse Data

Bayesian approaches have been used in situations that involve insufficient observations. Several studies have employed this approach for CR item analyses in various missing data designs (Cap, et al., 2010; Patz & Junker, 1999; Lee & Song, 2004; van Onna, 2002). In the Bayesian context, model parameters are assumed to be random and have *prior* distributions that reflect the uncertainty about the true values of the parameters. Hence, in theory, model identification is not an issue for estimating parameters (Jackman, 2009). However, non-identified (or weakly identified) models still raise several issues even in a Bayesian context and so Bayesian approaches to resolve identification issues have been discussed in previous literature (e.g., Jackman, 2009; Kass, et al., 1998).

In this section, the fundamental concept and the most popular computational methods of Bayesian inference are introduced, and then a Bayesian technique that uses informative priors for insufficient observations is reviewed.

II.4.1. Bayes' Theorem and Bayesian Inference

The fundamental idea of Bayesian inference is based on Bayes'(1763) theorem, which defines the law of conditional probability such that,

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}, \tag{II.4.1}$$

where:

$p(\boldsymbol{\theta} \mid y)$ is the *posterior* distribution, i.e., the conditional distribution of a set of unknown parameters, $\boldsymbol{\theta}$, given the data y,

$p(y \mid \boldsymbol{\theta})$ is the *likelihood* function, that is, a function of parameter $\boldsymbol{\theta}$ given the observed data y, and

$p(\boldsymbol{\theta})$ is the *prior* distribution, which reflects the researcher's *a priori* beliefs about the parameter, $\boldsymbol{\theta}$.

Note that the term $p(y)$ contains only the sample information (free of $\boldsymbol{\theta}$) and serves as a normalizing constant that ensures the posterior distribution integrates to 1. Considering this constant term, Equation II.4.1 can be formulated as,

$$p(\boldsymbol{\theta}|\, y) \propto p(y|\boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

where the symbol "∝" stands for "is proportional to,". The above can be written as,

$$\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}.$$

This expression summarizes the technical core of Bayesian inference. Based on this expression, two modeling stages can be constructed: (1) the specification of a *prior* distribution on parameters and (2) the specification of a model linking the data and parameters to build a *likelihood* function. Then, a posterior distribution is constructed based on the prior information combined with the information from the data, via the likelihood function.

This first stage is what differentiates Bayesian approaches from other analytic approaches. For instance, LC-SDT model specifications correspond to the second modeling stage, and the model parameters can be estimated by maximizing the likelihood function without consideration of the first modeling stage, that is, the priors. However, maximum likelihood estimation (MLE) has several limitations that might be addressed by a Bayesian approach that incorporates prior distributions in the model (e.g., Gelman et al. 2004; Jackman, 2009; Levy, 2009). For example, a Bayesian version of an LC-SDT model can be defined by assuming appropriate priors for the model parameters, which might alleviate identification issues in single rater designs. A more detailed review of the use of priors will be given in later sections.

II.4.2. Bayesian Estimation

After constructing a posterior distribution, different types of point estimators can be used within a Bayesian framework. One is to use the posterior mode (i.e., the maximum of the

posterior distribution) and a second is to use the posterior mean (i.e., the mean of the posterior distribution).

Posterior Mode Estimation (PME)

Similar to applying maximum likelihood estimation (MLE), PME obtains the *maximum* of the posterior distribution to obtain point estimates of the parameters. Specifically, the maximum of the log posterior function is found in PME, where the log posterior function is the log likelihood function *plus* a log prior function. By considering prior information about parameters, PME is more useful than MLE and can be easily implemented in currently available software (e.g., Latent Gold) without intensive computational time. It has also been employed in previous research on LC-SDT models and has led to adequate recovery of the population parameters and moderate standard errors (DeCarlo et al., 2010; Kim, 2009; Park, 2011). An advantage of PME over a fully Bayesian approach via MCMC is that only the mode of the posterior distribution needs to be obtained, rather than sampling from the full posterior distribution.

Markov chain Monte Carlo (MCMC) Simulation

In Bayesian contexts, parameters can be estimated by using the *mean* of the posterior distributions. It is often difficult to obtain the posterior mean when the integral of the posterior distribution is intractable (i.e., too complicated to obtain an analytic derivation of the posterior distribution). Hence, simulation methods can be used to compute the posterior mean and other statistics given a set of random samples from the marginal posterior distributions. This simulation approach has an advantage over analytic approaches (i.e., MLE or PME), in that it is flexible for complex modeling of various missing data structures, including single rater designs.

MCMC is one of the most popular simulation methods and refers to a general method of drawing values of parameters (i.e., $\boldsymbol{\theta}$) from approximate distributions and then correcting those draws to approximate the target posterior distribution (Gelman, et al, 2004). The samples are drawn sequentially, with the distribution of the sampled values depending on the last value drawn; hence, the draws form a Markov chain. Then, a large number of random sequences (generated by the Markov chain) will generally constitute a sample from the posterior distribution based on the Monte Carlo principle, which states that "anything about a random variable $\boldsymbol{\theta}$ can be learned by sampling many times from $f(\boldsymbol{\theta})$, the density of $\boldsymbol{\theta}$" (Jackman, 2009, p.133).

Due to the iterative nature of an MCMC algorithm, researchers have (e.g., Gelman et al. 2004; Jackman, 2009) recommended investigating whether enough iterations (or samplings) have been drawn to reach the target posterior distribution with sufficient accuracy. First, in order to use MCMC, one needs to verify if the sampler has converged to stationarity (representing the target distribution) after an initial *burn-in* period (i.e., a finite number of iterations designed to remove dependence from the starting location). The most popular way to diagnose convergence is to plot the iterative history of the parameters from the simulation runs and to monitor their trends, with plots often called 'time-series plots', or 'trace plots' (Fox, 2010; Jackman, 2009).

Once convergence has been reached, an additional large number of iterations are needed to obtain samples for posterior inference (Spiegelhalter et al., 2011), and so one needs to examine if the length of the sampler is long enough to recover estimates with sufficient accuracy. The accuracy of the posterior estimates is often reviewed by the Monte Carlo (MC) error for each parameter. MC error is an estimate of the difference between the mean of the sampled values and the true posterior mean. Spiegelhalter et al. (2011) suggested that the simulation

should be run until the MC error is less than about 5% of the sample standard deviation, as a rule of thumb.

II.4.3. Using Priors for Model Identification

With regards to identification issues, Vermunt and Magidson (2005) noticed that the use of priors allows a model to be identified that would otherwise not be identified. In these cases, the prior information is just enough to uniquely determine the parameter values. Among the available techniques in Bayesian contexts, techniques relevant to priors that have been recommended for identification issues are reviewed in the following section: informative priors and Bayes' constants.

Informative Priors

As discussed in section II.4.1, the specification of priors is crucial in Bayesian inference. Two basic interpretations can be given for prior distributions (Gelman, et al., 2004, p.39): a *population* interpretation and *subjective state of knowledge* interpretation. The former indicates that the prior represents a population of possible parameter values, and the latter indicates that the prior reflects substantive information (and uncertainty) about parameters, assuming a sampled value can be considered as a random realization from the prior distribution.

Among the available prior distributions that satisfy the former role (e.g., covering all possible parameter values), two types of priors can be used. One can employ priors that involve no subjective knowledge (i.e., *non-informative* priors; Spiegelhalter, Thomas, Best, & Gilks, 1996). For example, if nothing is known about the parameter, then a non-informative prior is often used, such as a rectangular distribution over the feasible set of values of the parameter, assuming equal probabilities for all values, i.e., uniform prior.

On the other hand, one can use priors that reflect substantive information about the parameter (i.e., *informative* priors). The implementation of informative priors can be based on literature reviews or explicitly from an earlier data analysis (Gelman, et al., 2004), and often alleviates identification issues (e.g., Galindo-Garre, et al., 2004; Kass, et al., 1998).

Nonetheless, it should be noted that the use of informative priors gives estimates that move towards the prior distributions. For example, Van Onna (2002) reported that when using informative priors, posterior means were closer to the true values when the informative prior distribution was closer to the population value (i.e., the value used to generate the data in simulations) as opposed to those where non-informative priors were used. However, when the informative prior distribution is far from the population value, posterior means from non-informative priors were closer to the population value than those from informative priors.

Using Normal Informative Priors

For general model parameters, e.g., regression coefficients, it is convenient to adopt a set of univariate normal priors assuming that no information about the dependencies between parameters is available (Gelman, et al., 2004). The effect of using normal priors with a mean of a certain value (e.g., 0) is that the parameter estimates are smoothed towards that value (e.g., 0). Meanwhile, similar to a uniform prior, non-informative priors can be approximated by taking a univariate normal distribution with a large variance (Gelman, et al., 2004). With univariate normal priors, informative priors can be applied by specifying smaller variances depending on the amount of prior information; the smaller the variance in normal priors, the more the posterior parameter estimates are smoothed towards the mean of that prior.

Normal priors have been employed for the rater criteria and detection parameters in previous literature on rater effects and SDT models (e.g., Cao, et al., 2010; Jackman, 2004; Lee

& Wagenmakers, 2010) and also for item difficulty and discrimination parameters in the IRT literature (e.g., Curtis, 2010; De Boeck, 2008; Fox, 2010). For item discrimination parameters, IRT models assume that the parameters are positively related to a latent trait θ, and this restriction is considered by using truncated normal priors (e.g., Curtis, 2010; Fox, 2010) or log transformations of item discrimination parameters (e.g., Levy, 2006).

For example, non-informative univariate normal priors for rater parameters can be written as,

$$d_j \sim N(0, 0.01), \text{ where } d \in [0, \infty],$$

$$c_{j1} \sim N(0, 0.01),$$

where ' ~ ' stands for 'distributed as' and N(mean, precision) indicates a normal distribution with parameters of 'mean' and 'precision'(i.e., 1/variance), where small values of precision in a given example are equivalent to a large variance, indicating non-informative priors. A precision value of 0.01 can be found in previous literature to represent 'uncertainty' over those parameters. For example, in the application of latent class models, Jackson (2004) used a variance of 10^2 for the normal priors for a logistic regression coefficient for a latent trait and for the thresholds for response categories. For a restriction of positive values for the detection parameter, d, a range of d in "$d \in$ [min., max.]" is needed for specifying a truncated normal distribution.

In studies that use informative priors, smaller variance (larger precision) values are employed since the variance in normal priors determines the amount of prior information (e.g., by decreasing the variance, one can increase the effect of smoothing toward the mean value). For example, smaller variance (or larger precision) values can be found in a study on Bayesian estimation of logit parameters with small samples (Galindo-Garre et al., 2004). The authors applied three types of informative priors with variances of 4, 10, and 25. They found that among

those three normal priors, the one with the smallest variance performed best and the one with the largest variance performed worst.

Priors for PME

For latent classes and the conditional response probabilities, which are part of the LC-SDT model specification, Dirichlet priors have been used in literature (Galindo-Garre, et al., 2004: Gelman et al., 2004; Schafer, 1997). Dirichlet priors are so-called conjugate priors, since the prior distribution has the same form as the posterior distribution; in this case, a Beta distribution. Conjugate priors are convenient and often facilitate the derivation of posterior distributions.

The use of Dirichlet priors in the context of PME has been suggested by several authors (Galindo-Garre & Vermunt, 2006; Schafer, 1997; Vermont & Magidson, 2005) to resolve boundary problems. Boundary problems occur when more than one parameter estimate is close to the boundary of the parameter space. In applications of LC-SDT models, for example, boundary problems occur when estimates of a latent class size are zero or unity, or when estimates of the rater detection parameters are excessively large or indeterminate. Boundary problems often occur in small or sparse samples with MLE and cause highly biased estimates of parameters and large standard errors.

To account for possible boundary problems when using LC-SDT models, PME and Dirichlet priors have been used in prior research (DeCarlo, 2008; 2010; DeCarlo, Kim, & Johnson, 2011; Park, 2011). The Dirichlet prior can be applied to the conditional response probabilities $\pi_{y|\eta}$, which is the probability of the observed response y given the true (latent) category η,

$$p(\pi_{y|\eta}) \propto \prod_{m=1}^{M} \prod_{k=1}^{K} p(y_j = k \mid \eta_j = m)^{\alpha-1},$$

Further, the parameter α can be defined as,

$$\alpha = 1 + \frac{B\hat{\pi}_{jk}}{M},$$

where B are the Bayes' constants, *M* is the number of latent categories, *K* is the number of response categories, and $\hat{\pi}_{jk}$ is the observed marginal proportion for Y_{jk} (for further details, see DeCarlo et al., 2011; Vermunt & Magidson, 2005).

Here, α is defined in a way so that a zero value for a Bayes' constant is equivalent to adding no observations to the data (and so there is no impact of the prior on the posterior), which is equivalent to using MLE. Bayes' constants determine the strength of the prior distribution on the posterior and can be interpreted as adding observations to the data for smoothing the parameter estimates away from the boundary. For instance, with a larger value of *B*, the rater detection parameters are smoothed toward zero, but there is only a small effect on the rater criteria estimates (for a discussion of this, see Clogg, Rubin, Schenker, Schultz, & Wiedman, 1991; DeCarlo et al., 2011). With respect to the latent class sizes, a larger value of *B* smooths the class size estimates towards equality. Previous simulation studies have employed a value of 1 for Bayes' constants *B* in order to avoid boundary problems in designs with two raters, and good recovery of the parameters was found (DeCarlo, 2010; Park, 2011).

Chapter III

METHODS

This chapter outlines the simulations as well as the analysis of real-world data. Section III.1 describes the simulation conditions and model estimation methods for the simulation studies. The first simulation study examines LC-SDT models in single rater designs and the second simulation study investigates the model in partial second rater designs. In section III.2, analysis of a real-world dataset is illustrated.

III.1. Simulation Studies

Simulation studies are conducted to investigate possible solutions for rater performance assessment in single rater designs. Three types of possible remedies are considered in order to alleviate identification issues in fitting LC-SDT models in single rater designs: 1) constraints on the model parameters 2) Bayesian informative priors and 3) partially selected 2nd rater's scores. The first two remedies are examined in the first simulations, where ten raters' ratings are simulated without additional second rater observations (e.g., each examinee's response is examined by only one rater). The third remedy (e.g., back-reading) is examined in the second set of simulations, where 10% or 30% of examinees have additional ratings from a second rater. The second set of simulations also investigate the utility of using Bayesian informative priors for back-reading data.

Simulation 1: Single Rater Designs

Simulation 1 exam the performance of Bayesian approaches employing LC-SDT models (e.g., Equation II.1.5 with the logistic density function) in single rater designs. To be specific, the LC-SDT model can be written as;

$$g\left[P(Y_{ij} \le k \mid \eta_i) \right] = c_{jk} - d_j \eta_i,$$

where g is the cumulative logit function, Y_{ij} is the response variable for a rater j (j=1,…J) assigning discrete score k (k=1, …, K response categories) to examinee i's (i=1, …, N) response on a CR item. Here, η is the true latent category for the constructed response (η=0,…, K–1), and c and d indicate rater criteria and rater detection parameters, respectively.

For the examination of identification issues in single rater designs, a preliminary simulation found that a typical LC-SDT model showed non-unique estimates (i.e., the model is not identified). These results are reported by statistical software (Latent Gold; Vermunt & Magidson, 2005) that examine the Jacobian matrix to determine identification. To alleviate the identification problem, Simulation 1 includes two remedies; 1) LC-SDT models are modified by using parameter restrictions and PME, or 2) highly informative normal priors for rater parameters are used in a fully Bayesian approach with MCMC.

Table III.1

Data Generation Conditions in Simulations

	Data generation conditions			
	2nd rater (%)	Rater detection (for 10 raters)	Examinee/Rater ($\bar{n}_j$)	Total sample size
Simulation 1				
	No (0%)	Constant d =2	100	1000
	No (0%)	Constant d =2	500	5000
	No (0%)	Varied d's (d=1,1,2,2,3,3,4,4,5,5)	100	1000
	No (0%)	Varied d's (d=1,1,2,2,3,3,4,4,5,5)	500	5000
Simulation 2				
	Yes (10%)	Varied d's (d=1,1,2,2,3,3,4,4,5,5)	100	1000
	Yes (10%)	Varied d's (d=1,1,2,2,3,3,4,4,5,5)	500	5000
	Yes (30%)	Varied d's (d=1,1,2,2,3,3,4,4,5,5)	100	1000
	Yes (30%)	Varied d's (d=1,1,2,2,3,3,4,4,5,5)	500	5000

Note. $\bar{n}_j$ indicates the average number of examinees per rater

Simulation conditions. In Simulation 1, four conditions with combinations of two conditions regarding d specification (e.g., constant d's vs. varied d's across the ten raters) and

sample sizes (e.g., average number of constructed responses per rater, $\bar{n}_j$=100 vs. $\bar{n}_j$=500) are compared in order to examine how sample sizes and the true rater detection values affect estimation results (see Table III.1 for details).

A previous simulation study (DeCarlo, 2005) with a small sample size of 125 with 5 response categories in a fully-crossed data set (e.g., all raters evaluate all examinees' responses) found that parameter recovery was marginal; hence, a sample size with a mean of 100 examinees for 10 raters (in total, N=1000) are used here. On the other hand, simulations with sample sizes of $\bar{n}_j$=500 (N=5000 for 10 rater) are suggested in part by DeCarlo's (2010) observation that 200 to 10,000 examinees' responses are typically obtained and scored by 10 to 80 raters on any given test day for many real-world large-scale tests.

The first situation considered is that all of the raters have an equal *d* value. A value of 2 is used for the population *d* value. Another situation considered is where different raters have different *d* values. For varied detection parameters, values of *d* used are 1, 1, 2, 2, 3, 3, 4, 4, 5, and 5 for Rater 1 to 10, which covers a range of detection from poor to excellent (for the logistic model) and are similar to values found in previous research (DeCarlo, 2008; 2010; DeCarlo et al., 2011).

Data generation. The simulations consist of 100 replications for each condition, as used in earlier simulation studies (e.g., DeCarlo, 2008). Data for the simulations consist of CR items with 6 response categories and were generated using a SAS macro for the LC-SDT model written by DeCarlo (2008). Similar to the original study, fully-crossed data (for ten raters) were generated based on the LC-SDT model; then single rater designs were created by generating missing values in the fully-crossed data.

The following population values for LC-SDT model parameters were used in data generation. First, latent class size values of (.08, .17, .25, .25, .17, .08) are chosen to approximate a normal distribution, which is consistent with previous simulation studies and results found in real-world applications (DeCarlo, 2005, 2008).

Table III.2

Population Values for the LC-SDT Model for the Equal d Condition

	Rater									
Parameter	1	2	3	4	5	6	7	8	9	10
d	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
c_1	−1.0	−1.0	0.0	0.0	1.0	1.0	2.0	2.0	3.0	3.0
c_2	1.0	1.0	2.0	2.0	3.0	3.0	4.0	4.0	5.0	5.0
c_3	3.0	3.0	4.0	4.0	5.0	5.0	6.0	6.0	7.0	7.0
c_4	5.0	5.0	6.0	6.0	7.0	7.0	8.0	8.0	9.0	9.0
c_5	7.0	7.0	8.0	8.0	9.0	9.0	10.0	10.0	11.0	11.0
shift	−2.0	−2.0	−1.0	−1.0	0.0	0.0	+1.0	+1.0	+2.0	+2.0
		Lenient			**No shift**				**Strict**	

Table III.3

Population Values for the LC-SDT Model for the Varied d Condition

	Rater									
Parameter	1	2	3	4	5	6	7	8	9	10
d	1.0	1.0	2.0	2.0	3.0	3.0	4.0	4.0	5.0	5.0
c_1	−1.5	2.5	−1.0	3.0	−0.5	3.5	0.0	4.0	0.5	4.5
c_2	−0.5	3.5	1.0	5.0	2.5	6.5	4.0	8.0	5.5	9.5
c_3	0.5	4.5	3.0	7.0	5.5	9.5	8.0	12.0	10.5	14.5
c_4	1.5	5.5	5.0	9.0	8.5	12.5	12.0	16.0	15.5	19.5
c_5	2.5	6.5	7.0	11.0	11.5	15.5	16.0	20.0	20.5	24.5
Shift	−2.0	+2.0	−2.0	+2.0	−2.0	+2.0	−2.0	+2.0	−2.0	+2.0

As in prior studies (DeCarlo, 2008, 2010; Park, 2011), raters' criteria values are initially located at the distribution mid-points. For example, a rater with a detection population value of 2 has criteria locations for the five categories at 1, 3, 5, 7, and 9, respectively. This assumes that the response criteria are located at the intersection points of adjacent distributions between adjacent latent classes; empirical studies have found that this assumption is reasonable (DeCarlo, 2008) and it is also optimal in certain situations.

Subsequently, rater's criteria values are shifted from these intersection points to represent situations with rater effects such as rater leniency and severity. For example, shifting the raters' criteria by a positive value indicates greater severity, whereas negatively shifted *c* values indicate greater leniency. In the constant *d* conditions, as shown in Table III.2, the criteria for the first two raters are shifted down from the intersection point locations by 2, and the criteria for the next two raters are shifted down by 1. The 5^{th} and 6^{th} raters' criteria remain at the intersection points. The criteria for the next two raters are shifted up from the intersection points by 1, and the criteria for the last two raters are shifted up by 2. This situation is guided by a previous simulation study (DeCarlo, 2008). In the varied *d* conditions (shown in Table III.3), the criteria for a specific *d* value are shifted down by 2 or up by 2.

Parameter constraints. A summary of restrictions used for each simulation study is shown in Table III.4.

Table III.4

Simulations using Parameter Constraints

	LC-SDT Models	Estimation	Priors
Simulation 1 (1 rater for each CR)	Constraints on d(=3) and $p(\eta)$ No constraint	PME Bayesian	Bayes' constants Normal priors
Simulation 2 (2 raters for some CRs)	No constraint No constraint	PME Bayesian	Bayes' constants Normal priors

In regards to using restrictions on model parameters, one can assign a specific value for all of the raters' detection, which is similar to the approach used for IRT model simplification (e.g., assigning a value of 1 for the item discrimination parameter). In Simulation 1, a value of 3 is assigned to the rater detection parameters, since previous literature (DeCarlo, 2008) on a large scale assessment found a mean of about 3.5 (*SD*=.9) for rater detection (for six response categories). Also, the simulations include conditions where the fixed values of *d* are incorrectly specified, and in particular, the true values for rater detection are 2 for all ten raters (in constant *d* conditions) or varied from 1 to 5 (in varied *d* conditions).

In addition to the rater detection parameters, one can set the latent class sizes (e.g., the probabilities of the latent classes). For example, with 6 response categories based on the normal distribution, the latent class sizes used are (.08, .17, .25, .25, .17, .08), which is similar to latent class sizes found in a large scale assessment (DeCarlo, 2005, 2008).

With these parameter constraints on rater detection (e.g., *d*) as well as latent class sizes (e.g., $p(\eta)$), one can estimate the rater criteria locations (e.g., *c*'s) which is useful for studying various rater effects. These model constraints allow a LC-SDT model to be just-identified in single rater designs; e.g., all the *c* parameters are unique. Hence, Simulation 1 apply this approach.

For the parameter constraint approach, PME with Bayes' constants for the latent classes and for the response categories (with Dirichlet priors) are used. Previous simulation studies employed Bayes' constants of 1 in order to avoid boundary problems, and found good recovery of parameters (DeCarlo, 2010; Park, 2011). Hence, Bayes' constants of 1 are considered here for the single rater designs.

Normal priors. Bayesian estimation with MCMC is used to examine the effect of using informative priors for *c* and *d*. Regarding the different levels of precision (of 0.25, 0.1, and 0.04) in normal priors, a simulation study on logit parameters with small samples found that a normal prior with a precision of 0.25 performs best and one with 0.04 performs worst (Galindo-Garre et al., 2004). Hence, in the current study, highly informative priors with precision are considered, such that,

$$d_j \sim N(3,\ 0.25), \text{ where } d \in [0, \infty],$$

$$c_{j1} \sim N(1.5, 0.25),$$

where plausible values of 3 for the detection and 1.5 (i.e., the midpoint) for the first criterion are used. For the rest of the criteria, the increments in criteria (e.g., c_{j2} to c_{j1}) are estimated in order to obtain criteria values, guided by other literature that employ MCMC for ordinal responses (e.g., Curtis, 2010; Jackman, 2009). Since the criteria increments and detection parameters have similar properties (e.g., representing the distance between response category distributions), the same priors for *d*'s are used for the *c* increments (e.g., $N(3,\ 0.25)$).

In the context of MCMC, Dirichlet priors via a gamma distribution method are used for the latent class sizes. Gamma distributions are often used to generate Dirichlet distributions based on the fact that if a random variable X_i is distributed as Gamma(α_i,1), then the vector $(X_1 / \sum(X), ..., X_n / \sum(X))$ follows a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_n$. The parameters in the gamma distribution, α's, are assumed to be distributed as $N(0,0.01)$; with a precision of 0.01, this prior can be considered as being non-informative (Jackson, 2004).

Simulation 2: Partial Second Rater Designs

Simulation 2 exams the performance of LC-SDT models in situations with partial second raters. In applications, partial second raters— as low as 1% for back-reading— allow the model

to be identified. Hence, the performance of the LC-SDT model in this sparse situation is examined in Simulation 2, without any constraints on the model parameters. The estimation of the LC-SDT model is examined by using PME with Bayes' constants. In addition, the utility of Bayesian informative priors (via MCMC) is examined (as in Simulation 1).

Simulation conditions. Four conditions with combinations of two conditions regarding the number of constructed responses per rater (e.g., $\bar{n}_j$=100, vs. $\bar{n}_j$=500) and the proportion of second raters (e.g., 10% vs. 30%) are simulated (see Table IV. 1 for details). Unlike Simulation 1, Simulation 2 only considers the condition with varied *d* values. For partial second rater designs, a proportion of second raters of 10% and of 30% are used, given that 10% of 2nd raters is considered to be the lowest value in operational practice of performance assessment (Jones & Vickers, 2011).

Data generation. Similar to Simulation 1, the simulations consists of 100 replications for each condition, with 6 response categories and 10 raters. The same population values for 10 raters in Simulation 1 are used (e.g., *d*=1, 1, 2, 2, 3, 3, 4, 4, 5, and 5, and shifted *c* as shown in Table IV. 3).

A SAS macro written by DeCarlo (2008) for the LC-SDT model was used to generate the simulated data. Analogous to Simulation 1, the fully-crossed data were generated first, and then data with partial second raters were created by generating missing values in the fully-crossed data according to the rater design.

In assigning two raters to each constructed response, previous literature employing LC-SDT approaches (DeCarlo, 2010; Park, 2011) have used two types of incomplete designs: the balanced incomplete block (BIB) design and the unbalanced design. A set of responses rated by each pair of raters is referred to as *blocks* and the term *unbalanced* indicates situations where

each pair of raters scores different numbers of responses rather than an equal number (i.e., *balanced*). Previous simulation studies (DeCarlo, 2010; Park, 2011) found similar patterns for parameter recovery between balanced and unbalanced incomplete designs; hence, only unbalanced incomplete designs are considered in this paper.

Table III.5

An Unbalanced Incomplete Design (10 rater pairs, N=1000, 10% linkage)

Rater	2	5	6	1	3	9	10	4	8	7	Total
	27										
	3	3									
		45									
		5	5								
			45								
			5	5							
				90							
				10	10						
					90						
					10	10					
						90					
						10	10				
							90				
							10	10			
								135			
								15	15		
									135		
									15	15	
										153	
	17									17	
Total/Rater	47	53	55	105	110	110	110	160	165	185	1100

Note. Numbers indicate the number of observations
(**bold**: rated by only one rater; non-bold: rated by two raters, i.e., linked)

Similar to simulations performed by DeCarlo (2010), ten rater pairs for ten raters are considered in Simulation 2, which is close to the minimum number of rater pairs (i.e., nine) needed in order to connect all the raters given the ten raters. Table III.5 shows an unbalanced

incomplete design that shows the number of constructed responses assigned to a particular rater (indicated by each column with different rater ID) in the case of $\bar{n}_j$ =100 with 10% 2nd rater scorings; for the cases of $\bar{n}_j$ =500, the number of examinees for each rater (n_j) is multiplied by 5 and for the cases of 30%, 2nd rater scorings n_j is multiplied by 3. Since the population values of *d* are ordered in increasing magnitude from Rater 1 to 10 (similar to previous simulation studies), the raters are randomly allocated to the 10 columns of the design, as shown in Table III.5, so that the sample sizes are not systematically related to the value of *d*.

Computational Methods

Among several software packages that can be used to fit the latent class SDT model, Latent Gold (version 4.5, Vermunt & Magidson, 2007) is used to obtain parameter estimates via PME with Bayes' constants and OpenBUGS (Open-source version of Bayesian inference Using Gibbs Sampling, version 3.2.1, Spiegelhalter et al., 2011) is used to examine univariate normal informative priors via MCMC.

Posterior mode estimation (PME). Latent Gold employs the expectation–maximization (EM) algorithm followed by the Newton-Raphson procedure to obtain posterior mode estimates (similar to what is done for MLE) of the parameters. 100 replications for each simulation condition were generated using a SAS macro written by DeCarlo (2008), which generates multiple simulated data sets and a DOS batch file, which allows analysis of those data sets via Latent Gold repeatedly.

One complication that has been recognized in the application of latent class analysis is *label switching* (DeCarlo, 2008; McLachlan & Peel, 2000); which is relevant to the coding of the latent categorical variable η being arbitrary, e.g., either the first or the last class can be assigned a value of zero. In other words, the posterior mode has the same value for the switched solution

($\eta=0, 1,\ldots, K-1$ vs. $\eta=K-1, K-2,\ldots,0$), however, the sign of d is reversed as is the order of the latent classes. Hence, a SAS macro (DeCarlo, 2008) is used to strip out and to summarize the Latent Gold output while it checks for and adjusts for label switching.

Another well-known problem in latent class analysis is that the solution could represent a local maximum not the global maximum. To avoid this problem, the number of starting values was increased from the default value of 10 to 20 as suggested in previous research (DeCarlo, 2008; Kim, 2009).

Bayesian estimation. In order to estimate parameters, OpenBUGS uses three types of MCMC algorithms (i.e., Gibbs, Metropolis Hasting and slice sampling) to sample values of the unknown parameters from their conditional (posterior) distribution given previously sampled values (i.e., Markov chains).

A SAS macro, OpenBUGSio (Smith & Richardson, 2007), was used to create a DOS batch file, which calls OpenBUGS and saves the sampled values and the output summary. Since OpenBUGSio is designed for one call, a SAS macro was written to call OpenBUGS repeatedly for replications (100 replications per condition, as also used for PME) and also to summarize the output. The SAS macro contains commands to strip out the simulated values for each replication, which can then be used to review the convergence via trace plots. In particular, the simulated values were reviewed by the R-package 'coda' (Plummer, Best, Cowles, & Vines, 2011), which is designed to analyze and diagnose MCMC simulation outputs.

For each replication, a 5000 iteration burn-in are followed by 30,000 updates with two chains (because, multiple chains reduce the variability and dependence on the initial values and are easier to establish convergence by comparing different chains). The number of burn-in and updates are determined by the consideration of the magnitudes of the MC errors and the trace

plots from several preliminary simulations, as suggested by Spiegelhalter et al. (2011). Given these numbers of iterations, the MC errors for the model parameters were less than about 5% of the sample standard deviation, and the trace plots with two chains indicated stability (except the simulation 1 with the smallest sample sizes showing a marginal stability). An example of trace plots and MC errors from one simulation is shown in Appendix E.

While LatentGold only takes a few seconds to estimate the model parameters, Openbugs requires much longer run times. Table III.6 summarizes the computational time for each condition using Openbugs. Openbugs took about an hour for the smaller samples ($\bar{n}_j$=100) and about four hours for the larger sample ($\bar{n}_j$=500).

Table III.6

Average Computational Time for Each Simulation Condition via Openbugs

	Data generation conditions			Computational time (in min.)	
	2nd rater (%)	Rater accuracy	Sample size ($\bar{n}_j$)	5000 Burn-in	30000 Sampling
Simulation 1					
	No (0%)	Constant *d*'s	100	14	57
	No (0%)	Constant *d*'s	500	54	217
	No (0%)	Varied *d*'s	100	20	58
	No (0%)	Varied *d*'s	500	55	222
Simulation 2					
	Yes (10%)	Varied *d*'s	100	15	60
	Yes (10%)	Varied *d*'s	500	61	243
	Yes (30%)	Varied *d*'s	100	20	90
	Yes (30%)	Varied *d*'s	500	67	269

Parameter Recovery

LC-SDT model parameters (e.g., rater parameters and latent class sizes), and their standard errors and classification accuracy (e.g., *Pc*) are examined to compare the performance of the different approaches.

Model parameters. Model parameters such as *c*, *d*, and *p*(η), are examined in terms of the bias, the (absolute) percent bias, and the mean squared error (MSE). Bias is calculated by taking the average of the parameter estimate minus the population value, and the percent bias is the absolute value of the bias divided by the population value times 100. The absolute value of percent bias is used to avoid confusion due to the sign of the bias. MSE is calculated as the average of the squared difference between the parameter estimate and the population value, which reflects both the variance of the estimator and its bias.

The model parameters are also reviewed graphically using plots that compare the true values and estimates of model parameters and plots that show relative criteria, which are useful to detect various rater effects. The notion of relative criteria was introduced by DeCarlo (2005) to compare the relative locations of the response criteria (i.e., relative to the highest and lowest underlying distributions), such as,

$$\text{rel } c_{jk} = \frac{c_{jk}}{(K-1)d_j}, \qquad \text{(III.1)}$$

where *K* is the number of latent classes. Applying this equation, the estimates of c_{jk} and d_j are used to compare all of a rater's utilization of criteria simultaneously. This approach has the convenient advantage that is allows one to compare the criteria locations across different values of *d*.

Standard errors/posterior standard deviations. Standard errors (in PME) and posterior standard deviations (in Bayesian estimation) for each model parameter are examined in terms of the standard deviation of the parameter estimates across the replications (that serve as the population value), which include the mean of the estimated standard errors/posterior standard deviations and their differences (e.g., bias).

This standard deviation, in turn, can be used to compute Monte Carlo standard error (MCSE) for replication, as the standard deviation across the replications divided by the square root of the number of replications. It should be noted that the generic term of MC error in MCMC simulation is computed as the standard deviation across the sampled values within each replication.

Classification accuracy. Classification accuracy is reviewed in terms of the proportion correct (*Pc*) and the adjusted proportion correct (λ) based on raw scores and on predicted latent classes compared to the true classes, which are known in simulations. In Simulation 2, raw scores from the two raters are averaged, and the classification accuracy for the average score is reviewed. In addition to PC and lambda, Pearson's correlation and Kendall's tau are also examined.

III.2. Empirical Study

The proposed approaches to LC-SDT models in single rater designs also are applied to real-world data: PIRLS (progress in international reading literacy study) 2006 reliability data sets (Mullis, et al., 2006). Since PIRLS contains international samples, only the US examinees are used to avoid the complex nature of sampling.

Table III.7

Description of Item 4 and Item 5 in PIRLS 2006 Reliability Data

	Item 4	Item 5
Kendall Tau-b	0.93 (*n*=235)	0.83 (*n*=198)
Number of examinees	1023	991
Number of raters	8	7
Second rater %	23%	20%

In this dataset, examinees answered one or two CR items out of seven CR items with four constructed response categories. There are a total of 2,021 examinees whose responses were

scored by a total of eight raters (ID: 1, 2, 3, 4, 5, 6, 7, and 9), with the raters scoring anywhere from 17 to 489 examinees with a mean of 225 examinees. The applied rater design is an unbalanced incomplete design with the average proportion of 2nd raters being 22.7%. In other words, 77.3% of responses are graded by only one rater (expressed as **bold** in Table III.8 & Table III.9). The inter-rater agreements, measured by Kendall's tau (τ) coefficient, are between 0.85 to 0.95, and the analysis includes the two CR items that have the lowest and highest tau: Item 5 and Item 4 respectively. The number of examinees who took Item 4 is 1023, and 23% of these answers are reviewed by another rater. For Item5, 991 examinees took the item and 20% of their answers are reviewed by second raters.

For the data analysis, parameter estimates (including rater parameters and latent class sizes) are obtained with LC-SDT models using both PME and Bayesian estimation (with informative priors). The empirical analysis compares these two approaches. Also examined are rater effects via estimates of the relative criteria locations. A sensitivity is also used to investigate the impact of using different Bayes' constants in PME and different variance in normal priors in Bayesian estimation.

Table III.8

Item 4, Rater Design in PIRLS (N=1023, 2nd rater=23%)

	Rater							
	1	2	3	4	5	6	7	9
	40							
	1	1						
	22		22					
	25				25			
	2					2		
	16						16	
		151						
		28	28					
		27			27			
		6				6		
		24					24	
			157					
			5	5				
			3		3			
			26			26		
				36				
				3	3			
				4			4	
					84			
					15	15		
					4		4	
						157		
						26	26	
							151	
								10
Total/Rater	106	237	241	48	161	232	225	10

Note. Numbers indicate the number of observations (**bold**: only one rater; non-bold: linked)

Table III.9

Item 5, Rater Design in PIRLS (N=1000, 2nd rater=20%)

	Rater						
	1	2	3	4	5	6	7
	66						
	4	4					
	17		17				
	3				3		
	11					11	
	3						3
		134					
		29	29				
		11			11		
		7				7	
		17					17
			207				
			5	5			
			7		7		
			27			27	
			16				16
				37			
				3	3		
				3			3
					111		
					11	11	
					13		13
						259	
						15	15
							180
Total/Rater	104	202	308	48	159	330	247

Note. Numbers indicate the number of observations (**bold**: only one rater; non-bold: linked)

Chapter IV

RESULTS

This chapter presents the results of the simulation studies as well as the analysis of real world data. Section IV.1 examines the LC-SDT models in Simulation 1, which used single rater designs. Section IV.2 shows the results of Simulation 2, which used the partial second rater designs. Finally, in Section IV.3, analyses of large scale assessment data are presented.

IV.1. Results for Simulation 1 (Single Rater Designs)

The simulations in this section present results that examine how well the LC-SDT model performs in single rater designs where only one rater scores each examinees' answer. The simulation includes four conditions with all combinations of two conditions regarding average examinees per rater (e.g., $\bar{n}_j$ =100, vs. $\bar{n}_j$ =500) and rater detection (d) values (e.g., constant d's of 2 vs. varied d's of 1, 1, 2, 2, 3, 3, 4, 4, and 5 for ten raters) as shown in Table II.1.

100 replication data sets were generated with 10 raters scoring a CR item with 6 latent categories, where each rater scores different numbers of examines (i.e., an unbalanced design). The generated criteria values are shifted from mid-point criteria locations to represent situations with rater effects (e.g., rater severity-leniency), as discussed above. In the condition of constant d's, the degree of shift in rater criteria are −2, −2, −1, −1, 0, 0, +1, +1, +2, +2 for Rater 1 to Rater 10 respectively. When the rater detections are varied for ten raters, the shifts in criteria are generated as −2 and +2 for each adjacent value of d.

Simulation 1 examines two remedies for LC-SDT model identification issues: 1) parameter restrictions are used for the rater detection parameters d and the latent class sizes $p(\eta)$) and 2) highly informative normal priors are used for rater parameters and Bayesian estimation is used. To be specific, for 1), the parameter restriction of d =3 for all raters and of $p(\eta)$ =

(.08, .17, .25, .25, .17, .08) are used, along with Bayes' constants of 1 and PME. For 2), highly informative priors are used as discussed above (e.g., normal priors with a mean of c_1=1.5 and d=3, both with a variance of 4).

To study the performance of these approaches, parameter recovery is examined by using the bias, the (absolute) percent bias, and the mean squared error (MSE). The following criteria are used for percent bias; percent bias values of less than 5% are usually considered as trivial, values between 5% and 10% as moderate, and values greater than 10% as substantial (Curran, West & Finch, 1996).

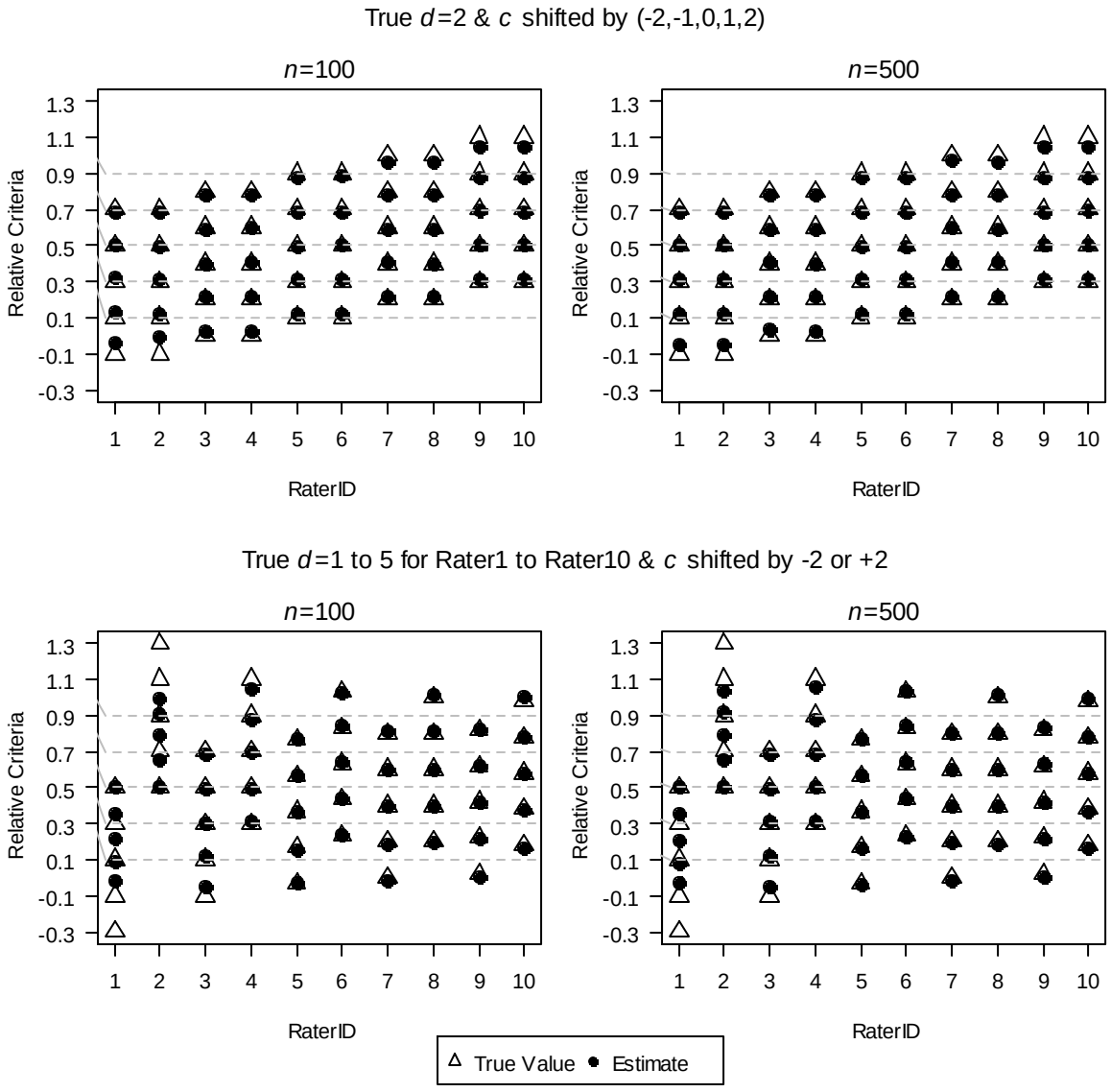
Parameter Estimates

Parameter constraints. Rater detection parameters and the latent class sizes are treated as known in this approach, hence only rater criteria parameters are estimated. Here, the true values of *d*'s are incorrectly specified in the fitted model to obtain some information about recovery in the presence of a (minor) misspecification of the parameter constraint. To be specific, the fitted model fixes all of the raters to have *d* values of 3, while the true values of *d*'s are either 2 for all raters (in a constant rater detection condition) or 1 to 5 (in a varied rater detection condition). Hence, the true values of rater criteria shown in Appendix A are for the rater detection value of 3. For instance, rater criteria of (-1, 1, 3, 5, 7) for a *d* of 2 is rescaled to (-1.5, 1.5, 4.5, 7.5, 10.5). More formally, the rescaling can be done by dividing the rater criteria values by the original rater detection value (e.g., 2) and multiplying the new rater detection value (e.g., 3).

The simulation outcomes (e.g., bias and MSE for the parameter recovery) are shown in Appendix A. Graphical comparisons of true values and parameters of relative criteria under different conditions are shown in Figure IV.1. The figure shows relative criteria values (on the y-axis) for each rater where the rater's ID is shown on the x-axis, and compares the true values

(shown in triangles) and the estimates (shown in filled circles) of the relative criteria in each simulation conditions.

Figure IV.1. Relative Criteria for PME with Bayes' Constants and Model Constraints (d=3) in Simulation 1



Note. The dotted lines at the mid-point locations indicate no shift in the rater criteria locations.

In the condition with the true detection parameters for all raters being constant (i.e., 2), most of the relative criteria estimates recover the criteria shifts. The patterns of rater criteria estimates are similar in both sample size conditions ($\bar{n}_j$=100 vs. $\bar{n}_j$=500) except that the sizes of MSE for the smaller sample size are much larger (2 to 4 times larger). MSEs for $\bar{n}_j$=100 are 0.17 to 2.18 and for $\bar{n}_j$=500 are 0.04 to 0.86. Substantial percent bias (greater than 20%) are only observed for rater criteria locations with negative or no shifts (for Rater ID= 1 to 6) for the first response category (e.g., c_{j1}) and for the largest negative shifts (i.e., −2 for Rater ID=1, 2) for the second response category (e.g., c_{12} and c_{22}).

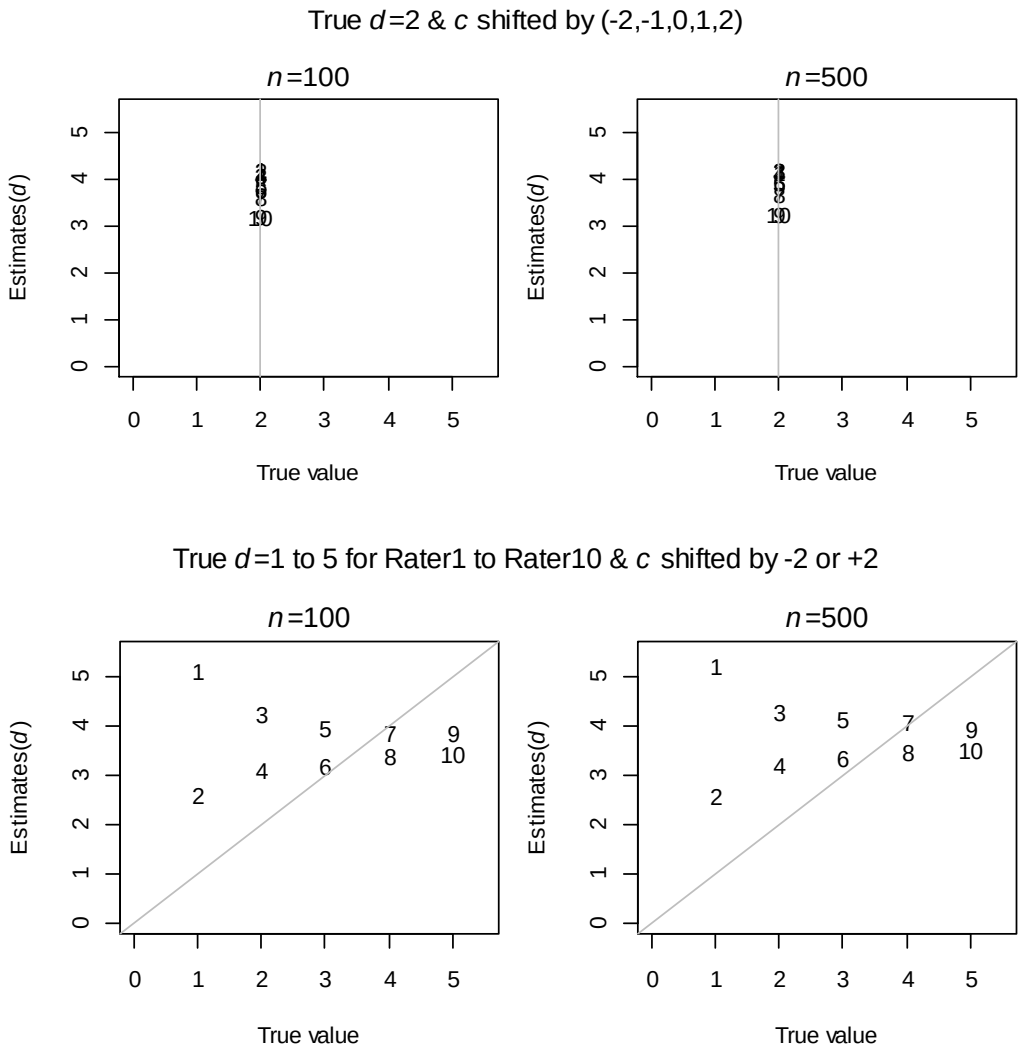
For the conditions with true detection parameters that vary for Rater 1 to Rater 10 (e.g., 1, 1, 2, 2, 3, 3, 4, 4, 5, and 5 respectively), the estimates of rater criteria show different patterns from the condition with the constant detection, especially for the raters with the lowest detection and a relatively large shift. For example, a broader range of MSE is observed as compared to the constant detection condition (e.g., MSEs ranged from 0.2 to 22.3 for $\bar{n}_j$=100 and from 0.03 to 16.3 for $\bar{n}_j$=500). The largest MSE values are for Rater 1 and Rater 2 who have the lowest value of detection (d=1). Also, shrinkage toward the middle categories for raters (with d=1) are substantial with a large percent bias, as shown in Figure IV.1.

Other than these exceptions, the shifts in criteria are overall very well recovered. Criteria estimates for Rater 3 and Rater 4 (with d=2: the same as the constant detection condition) show similar patterns to estimates under the constant detection condition, showing substantial shrinkage (greater than 20%) toward to the middle when negative shifts are present (e.g., c_{31} and c_{32}). Interestingly, Rater 9 and Rater 10 (with d=5: the highest detection value) criteria estimates show substantially negative bias for their first response criteria (e.g., c_{91} and c_{101}), which is the

opposite direction of bias in criteria for raters with low detection values (e.g., 1 and 2); this is most likely due to the shrinkage (of *d* towards zero, which also affects *c*) induced by PME.

Bayesian estimation with informative priors. There are no parameter constraints in the LC-SDT model in this approach, hence rater detection parameters, rater criteria parameters, and latent class sizes all are estimated and are presented in Appendix B1. This section starts with a review of the estimates of the rater detection parameters, then estimates of rater criteria and latent class sizes are reviewed.

Figure IV.2. Rater Detection Parameters for Bayesian Estimation with Normal Priors



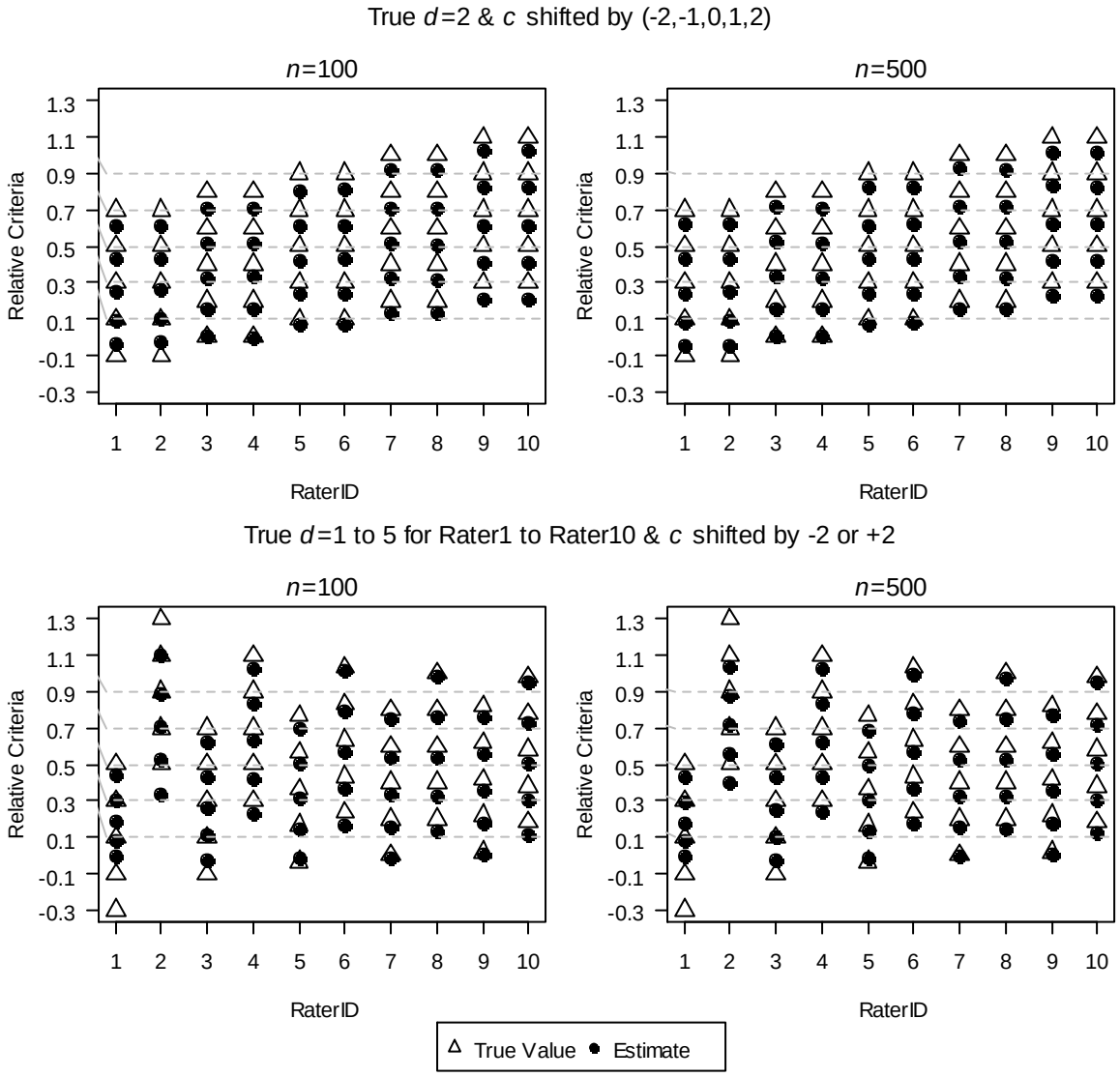
Note. Numbers (1 to 10) indicate the rater ID's.

Rater detection parameter estimates obtained for Bayesian estimation with informative priors in single rater designs are shown in Figure IV.2. Figure IV.2 compares the population values (on the x-axis) and the estimates (on the y-axis) of the rater detection parameters. The numbers, 1 to 10, on the plots indicate the rater ID's and the grey straight line provides a reference for cases where the true value and the estimate of a parameter are identical. The estimates of *d* ranged from around 3.2 to 4.3 (MSE: 1.6 to 5.2) under the constant *d* condition—where all the true detection values are 2—and 3.4 to 5.3 (MSE: 0.05 to 18.2) under the varied *d* condition— where the true detection values ranged from 1 to 5 (see Appendix B1 for details).

Regardless of the values of the *d*'s, all of the rater detection estimates are substantially biased and the magnitudes of the bias are large with a few exceptions: Rater 7, with a true detection value of d=4, which is close to the mean value of *d*, and where the criteria were negatively shifted by the greatest amount (i.e., −2), and Rater 6, with the true detection value of d=3, again close to the mean, and criteria also positively shifted by the greatest amount (i.e., +2).

For the varied detection condition, the detection estimates for raters with odd number ID's (e.g., 1, 3, 5, 7, and 9), who have negatively shifted rater criteria, are higher than their counterparts (i.e., Rater 2, Rater 4, …Rater 10, with the same detection values) who have positively shifted rater criteria. Interestingly, the larger the criteria shift compared to the detection value, the larger the bias in the corresponding detection estimates. For example, Rater 1 and Rater 2, who had true detection values of 1 and shifts of 2, which is twice as large as their detection values, show the largest difference between the detection estimates.

Figure IV.3. Relative Criteria for Bayesian Estimation with Normal Priors in Simulation 1

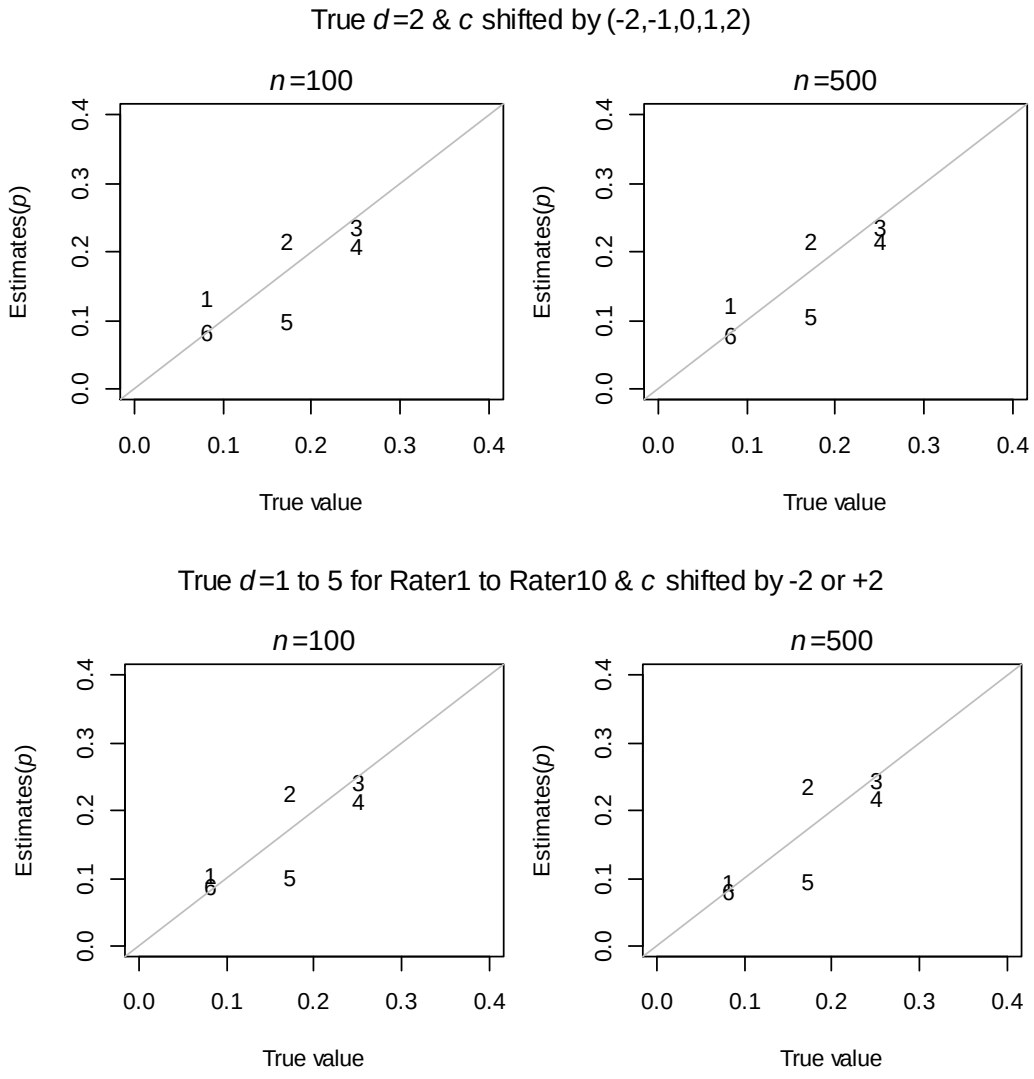


Note. The dotted lines at the mid-point locations indicate no shift in rater criteria.

Similar to the rater detection parameter results, the rater criteria estimates are substantially and positively biased with a few exceptions. The relative criteria estimates, shown in Figure IV.3, are negatively biased (i.e., under-estimation) but reveal the criteria shifts in most cases. It is noticeable that most of the relative criteria are under-estimated. In comparison to the parameter constraint approach from the previous section, MSE's for the rater criteria are much

larger—within a range of 0.17 to 82.19 under the constant *d* condition and 0.31 to 82.19 under the varied *d* condition. Similar to the results for parameter constraints, the estimates of the relative criteria recover the shifts in criteria, and the rater criteria estimates show shrinkage toward the middle, especially for the raters with low detection values (e.g, 1 and 2) and negative shifts in rater criteria.

Figure IV.4. Latent Class Sizes for Bayesian Estimation with Normal Priors in Simulation 1



Note. Numbers (1 to 5) indicates the latent classes

The estimates of the latent class sizes show similar patterns regardless of the conditions (e.g., constant or varied *d* values, and different sample sizes). Substantial biases are found except for the latent class categories of 3 and 6; there is positive bias at the lower categories (k=1 and 2) but negative bias at the higher categories (k=4 and 5), as shown in Figure IV.4 above. The magnitudes of the bias are within a range of −0.07 to 0.07 (see Appendix B for details), which are similar to those found in previous studies with two raters and PME (DeCarlo, 2008, 2010).

Standard Errors/Posterior Standard Deviations

Parameter constraints. Appendix A2 presents tables that show the estimates of the standard errors for the rater criteria in a single rater design. Since this approach constrains the detection parameters and the latent class sizes in the LC-SDT model, only the rater criteria are estimated. For the approach of parameter constraints via PME, the standard error estimates are computed using the inverse of the observed information matrix (see Vermunt & Magidson, 2005 for details). In the Appendix tables, the standard deviation of the parameter estimates across the 100 replications, SD, serves as the population value and the mean of estimated standard errors are named Mean SE and treated as the estimates of the SE's.

The biases are generally small to moderate (10 to 15% or less) which indicate that the standard errors are reasonably well estimated when the number of average examinees per rater ($\bar{n}_j$) is 500. However, when $\bar{n}_j$=100, there are a few cases with large bias (greater than 20%), especially for the raters who scored a small number of constructed responses (e.g., 30 and 50).

The Appendix also contains information about the Monte Carlo standard error (MCSE), indicated as SD in the table, which is computed as the standard deviation across the replications divided by the square root of the number of replications. The MCSE provides a guideline to assess the number of replications in simulations (Koehler, Brown, & Haneuse, 2009). For

example, a Monte Carlo 95% confidential interval for c_{11} can computed as (−0.70, −0.84) with a point estimate of −0.771 and an MCSE of 0.034. The interval does not include the true value of c_{11}, −1.5, which indicates that the number of replications, 100, is sufficient to detect significant bias.

Bayesian estimation with informative priors. Without any model constraints, the Bayesian approach provides estimates of the posterior standard deviation for all of the parameters, that is, the latent class sizes, rater detection, and the rater criteria locations. As shown in Appendix A2, the biases of almost all of the parameters are quite large (greater than 20%) and positive, which suggests that the SDs are over-estimated. Note that the applied models for Bayesian estimation and the PME are not equivalent (i.e., PME with the model parameter constraints); hence any comparison between two estimation methods requires caution.

Classification

Classification accuracy (e.g., proportion correctly classified) for the four conditions are shown in Table IV.1. LatentGold software (which is used for PME) reports the estimates of the proportion correct and lambda while the Bayesian approach requires extra programming steps to compute these values. Also, since the classification accuracy depends on the quality of the parameter estimates (as shown in Section II.1.2), one can argue that it is somewhat redundant to review the classification accuracy for MCMC; hence, only classification accuracy for the PME approach is reviewed here.

Proportion correct (PC_{pred}) is the estimated proportion of cases that are correctly classified and is obtained from the posterior probabilities. While PC_{pred} is available for both simulated and real world data, PC_{obt} is only available in a simulation, because, PC_{obt} is the obtained (not estimated) proportion of cases that were actually correctly classified in the

simulation (since the true latent class for each case is only known in a simulation). In addition, the results include PC_{raw}, which is the obtained proportion of cases that were correctly classified based on the average score. Another classification accuracy index, lambda, which adjusts the proportion correct using the largest latent class size, is also shown in Table IV.1. Two other measures of association between the model based classifications and the true latent classes are also shown in Table IV.1: the Pearson correlation *r* and Kendall's tau-b (with subscript *obt*).

Table IV.1.

Estimated and Obtained Proportion Correct and Correlations with True Latent Classes in a Single Rater Design (Simulation 1)

Number of Examinees ($\bar{n}_j$)	PC_{pred}	PC_{obt}	PC_{raw}	λ_{pred}	λ_{obt}	r_{obt}	τ_{b_obt}
Constant detection (d=2)							
100	0.585	0.478	0.392	0.447	0.294	0.794	0.707
500	0.590	0.483	0.389	0.453	0.306	0.797	0.710
Varied detection (d=1 to 5)							
100	0.567	0.512	0.432	0.424	0.342	0.815	0.734
500	0.564	0.528	0.431	0.418	0.368	0.819	0.740

Note. PC_{pred} = estimated proportion correct; PC_{obt} = obtained (in the simulation) proportion correct; PC_{raw} = obtained (in the simulation) proportion correct based on the raw score; λ_{pred} and λ_{obt} are the estimated and the obtained lambda; r_{obt} and $\tau_{b\text{-}obt}$ are the obtained Pearson correlation and tau-b.

As shown in the table, the estimated proportion correctly classified, PC_{pred}, tends to overestimate the proportion correctly classified in the simulation (i.e., PC_{obt}). The overestimation is moderate, generally around 0.04 to 0.11. Overestimation of PC_{obt} by PC_{pred} has been reported in similar simulation studies (DeCarlo, 2008, 2010), where it was also shown that the

overestimation is larger for smaller sample sizes. Compared to the PC based on the raw score, both PC_{pred} and PC_{obt} are about 14% to 20% higher in all conditions.

The magnitudes of PC and lambda found in the current study (with only one rater per examinee) are similar, but smaller, than those found in previous studies (with two raters per constructed response). For instance, DeCarlo (2008) reported a PC_{pred} of 0.623 and λ_{pred} of 0.478, whereas the current study reports PC_{pred} of 0.59 and λ_{pred} of 0.453 for the condition with constant *d*=2. In the same condition, Pearson correlation and tau-b were 0.866 and 0.792 in DeCarlo (2008), while these are 0.797 and 0.710 in the current study. The smaller values are expected because PC and the other statistics are partly a function of the number of raters per constructed response, and only one rater was used here.

With respect to the conditions where *d* varied, *obtained* classification accuracy indices and correlation statistics are higher than those for the constant detection condition. This may occur because the average of the true detection values (i.e., 3) in the varied detection condition is higher than the true detection values (i.e., 2) in the constant detection condition.

IV.2. Results for Simulation 2 (Partial Second Rater Designs)

Simulation 2 presents results that examine how well the LC-SDT model performs in single rater designs with partial second raters, where some examinees have additional ratings from a second rater. The simulation conditions include four conditions with combinations of two conditions regarding the average examinees per rater ($\bar{n}_j$ =100, vs. $\bar{n}_j$ =500) and the proportion of 2nd raters (e.g., 10% vs. 30%), as shown in Table II.1.

As a reminder, the data were generated for 10 raters rating a CR item with 6 response categories, where each rater scores different numbers of examines. The rater discriminations are generated with values of (1, 1, 2, 2, 3, 3, 4, 4, 5, 5) for Rater 1 thru Rater 10 respectively, and the

generated rater's criteria values are shifted from mid-point criteria locations, where the shifts in criteria are generated as −2 or +2 for each value of *d* (these values are identical to the varied *d*'s condition in Simulation 1).

Simulation 2 examines two approaches: 1) PME with Bayes' constants (with no parameter constraints) or 2) using informative normal priors in Bayesian estimation. To examine the performance of these approaches, parameter recovery and standard errors or posterior standard deviations are examined by using the bias, the (absolute) percent bias, and the mean squared error (MSE).

Parameter Estimates

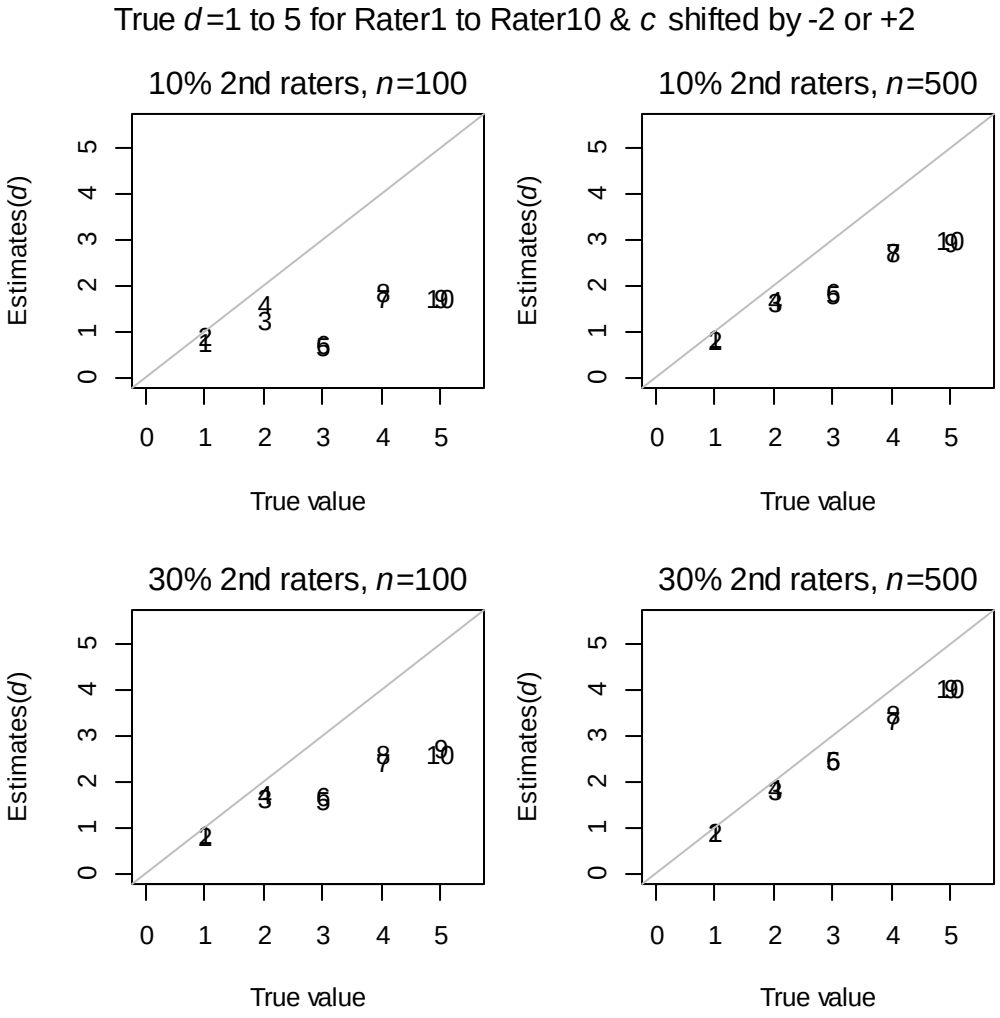
PME. Appendix C1 shows the simulation results for PME with Bayes' constants of 1 (for the response and the latent categories) to the situation with back-readings. This section presents results for the rater parameters first and then discusses the latent class sizes.

With respect to the rater detection parameters, all of the estimates are negatively biased, which indicates under-estimation (for details see Appendix C1). While the true values are between 1 and 5, the estimates are within ranges of 0.75 to 1.92 (MSE of 0.22 to 10.52) and 0.88 to 3.03 (MSE of 0.07 to 4.04) in the condition with 10% second raters for $\bar{n}_j$=100 and $\bar{n}_j$=500, respectively. For the condition with 30% second raters, the ranges are 0.90 to 2.78 (MSE of 0.09 to 5.64) for $\bar{n}_j$=100 and 0.96 to 4.10 (MSE of 0.02 to 0.97) for $\bar{n}_j$=500. This shrinkage towards zero is expected because it is a consequence of the use of PME (Vermunt & Magidson, 2005).

Figure IV.5 shows that under-estimation decreases as the sample size (i.e., the proportion of second ratings and the number of constructed responses per rater) increases. The magnitudes of bias in the condition with the largest sample size (i.e., 30% second rater with $\bar{n}_j$=500) are small (less than 20%); however, the bias in other conditions is substantially large (greater than

20%) especially for raters with population detection values higher than 3 (Rater 5 thru Rater 10). The magnitudes of MSE increase as the true values of *d* increase or the sample size decreases, which is consistent with results found in previous simulation studies with two raters (e.g., DeCarlo, 2010).

Figure IV.5. Rater Detection Parameters for PME in Simulation 2

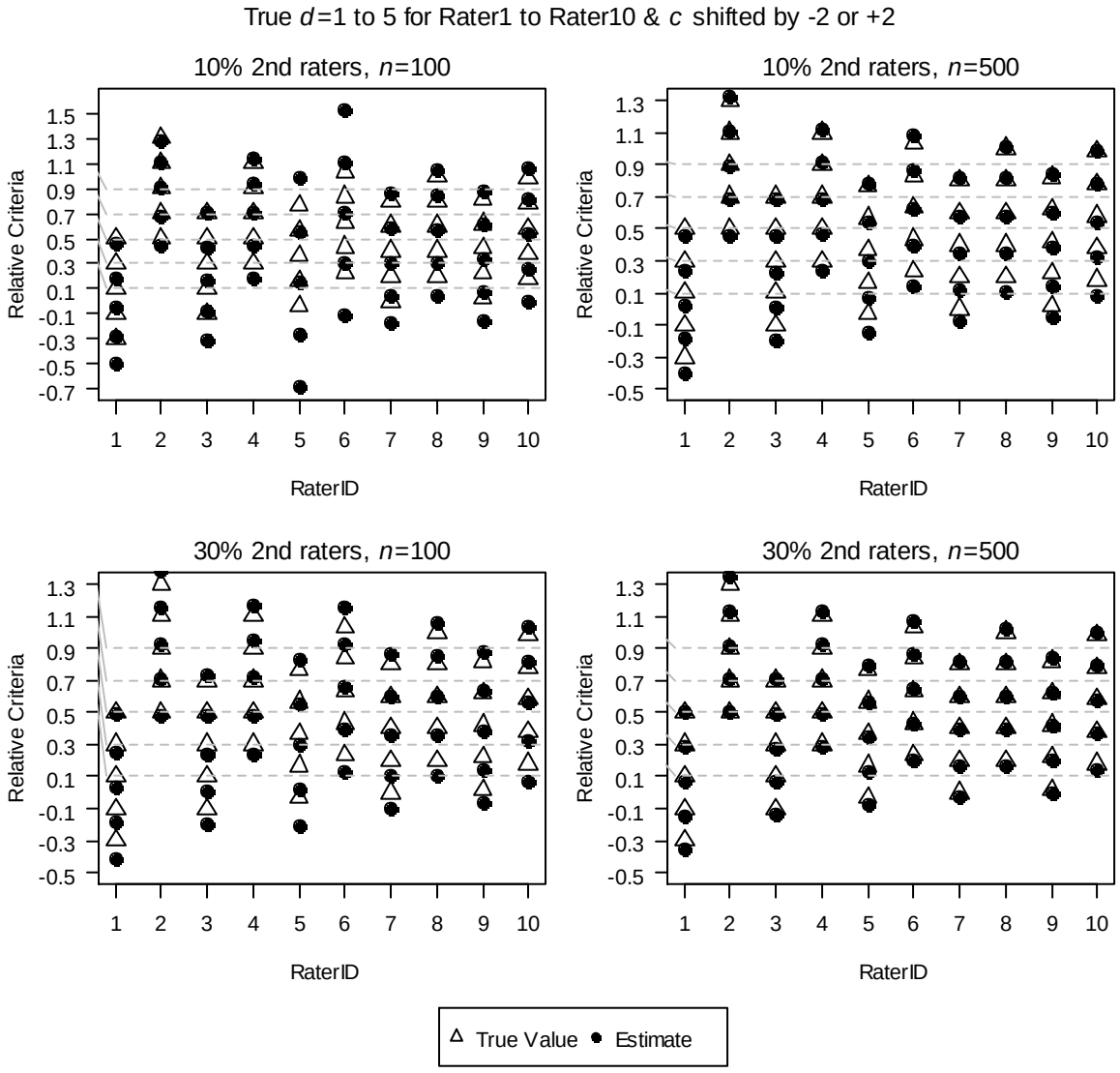


Note. Numbers (1 to 10) indicate the rater ID's

With respect to the rater criteria parameters, the relative criteria (shown in Figure IV.6) generally capture the shifts in the rater criteria, except for the smallest sample size (i.e., 10%

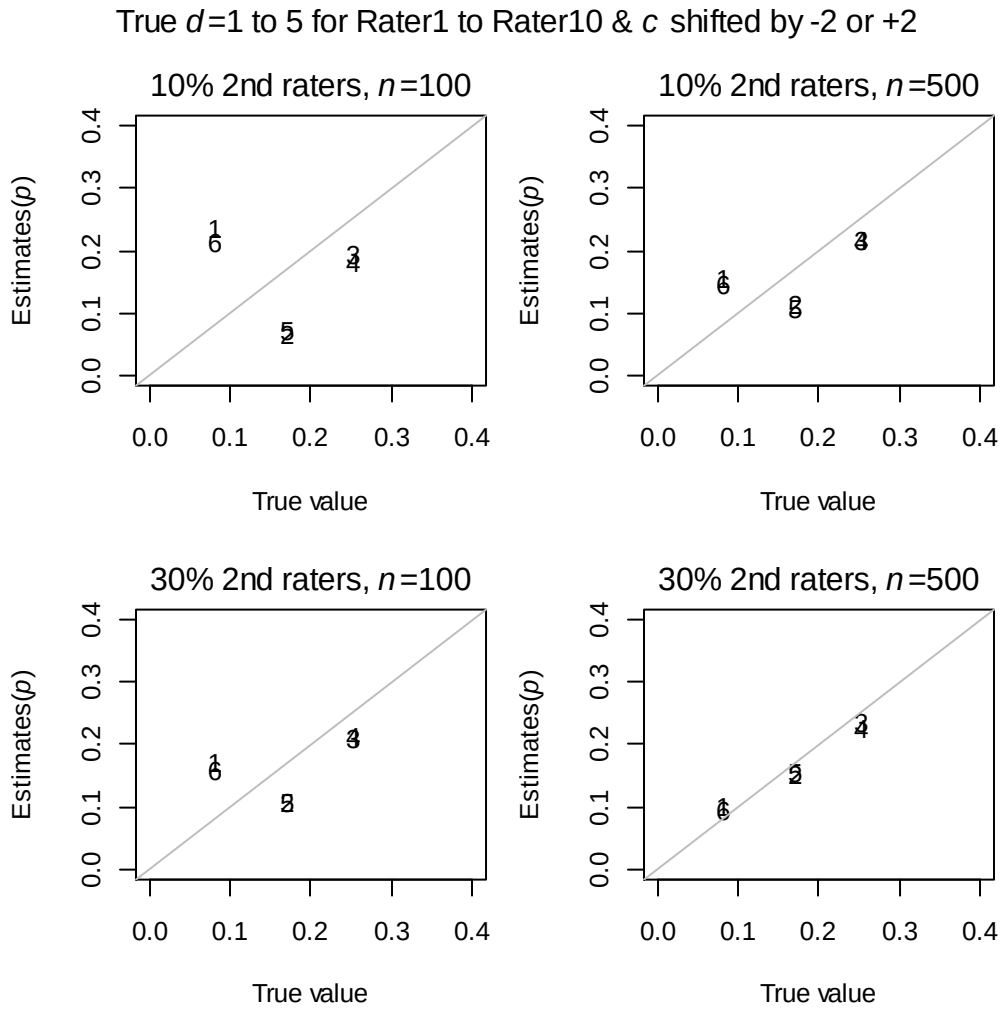
second rater with $\bar{n}_j$ =100). All of the estimates are negatively biased. In the condition with the largest sample size (i.e., 30% second rater with $\bar{n}_j$ =500), the biases in rater criteria estimates are small with some exceptions. On the other hand, in the other conditions with smaller sample sizes, most of the rater criteria estimates are substantially biased, except for Rater 2 and Rater 4, who have positive criteria shifts and the lowest detection values (e.g., 1 and 2).

Figure IV.6. Relative Criteria Estimates for PME in Simulation 2



Note. The dotted lines at the mid-point locations indicate no shift in the rater criteria.

Figure IV.7. Latent Class Sizes for PME in Simulation 2



Note. Numbers (1 to 5) indicates the latent classes

The estimates of the latent class sizes, shown above in Figure IV.7, show a fair effect of the sample size. While bias for all of the latent class sizes in the condition with the smallest sample size (i.e., 10% second rater with $\bar{n}_j$=100) are substantial, only bias for the end categories (e.g., *k*=1 and 6), which have the smallest class sizes, are substantial in all of the conditions. The bias in the other conditions is consistent with results found in previous simulations with two raters (e.g., DeCarlo 2008, 2010), where the smallest latent-class sizes tended to be

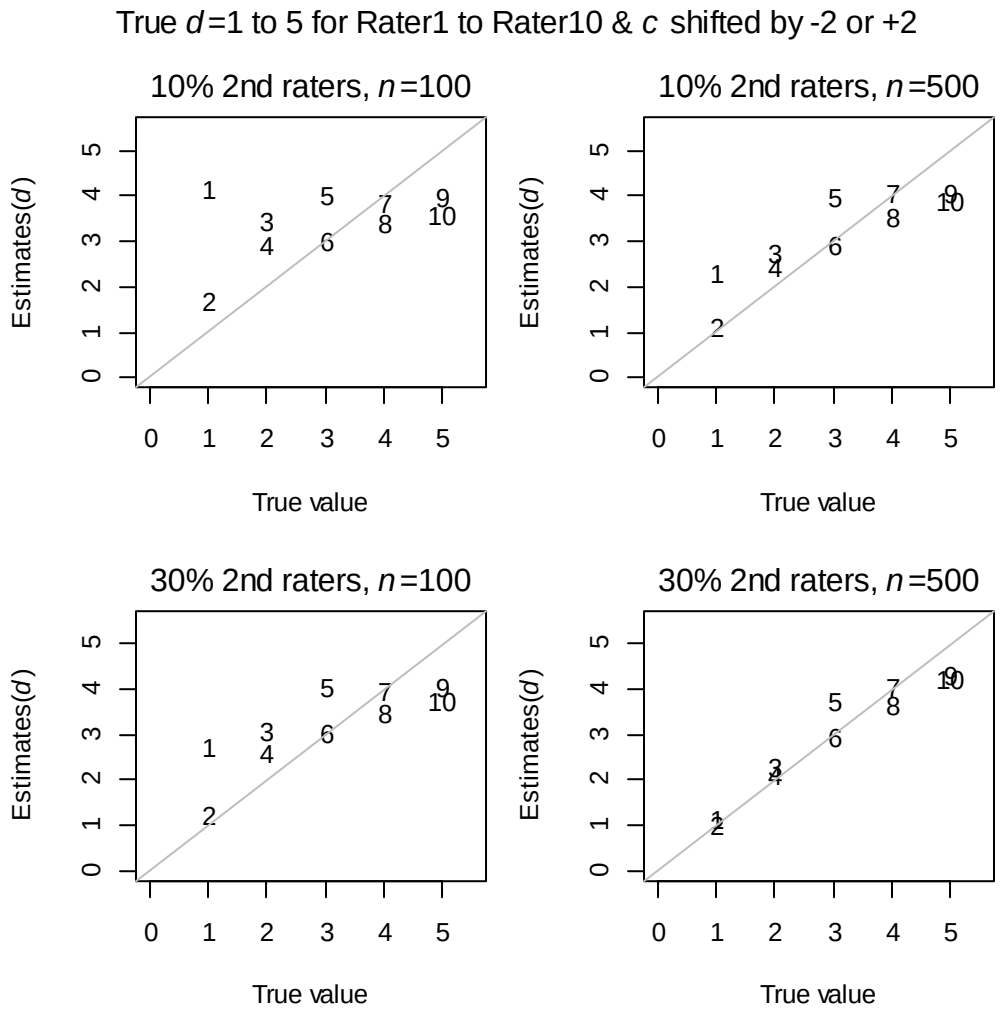
overestimated. With Bayes' constants (via PME), the latent class sizes show shrinkage toward a uniform distribution, where the end class sizes are over-estimated and the middle class sizes are under-estimated.

Bayesian estimation with informative priors. Appendix D1 shows the simulation results for Bayesian estimation with informative normal priors (with means of 1.5 for c_1 and 3 for d and variances of 4) for the situations with back-readings. Analogous to the previous section, this section presents results for the rater parameters first and then discusses the latent class sizes.

Regarding the rater detection parameters, the estimates tend to be positively biased when the true detections are low (e.g., d=1 to 3), but negatively biased when the true detections are high (e.g., d=4 to 5) (for details see Appendix D1); this again results from shrinkage to the mean of the prior (i.e., 3) and is commonly observed in the application of Bayesian methods. The detection estimates are within ranges of 1.75 to 4.20 (MSE of 0.11 to 10.98) and of 1.20 to 4.14 (MSE of 0.13 to 3.28) in the condition with 10% second raters at $\bar{n}_j$=100 and $\bar{n}_j$=500 respectively. In the condition with 30% second raters, the ranges of estimates are 1.28 to 4.13 (MSE of 0.11 to 4.69) at $\bar{n}_j$=100 and 1.07 to 4.36 (MSE of 0.03 to 0.77) at $\bar{n}_j$=500.

Figure IV.8 shows that the biases in rater detection decrease as the sample size (e.g., second rater proportion and the number of average examinees per rater) increase. The figure also demonstrates that, analogous to the simulations *without* back-readings (for the varied detection condition), the detection estimates for raters with odd number ID's (e.g., 1, 3, 5, 7, and 9), who have negatively shifted rater criteria, are higher than their counter-partners (i.e., 2, 4, 6, 8, and 10) with positively shifted rater criteria (given the same detection value), except for the condition with the largest sample size.

Figure IV.8. Rater Detection Parameters for Bayesian Estimation with Normal Priors in Simulation 2

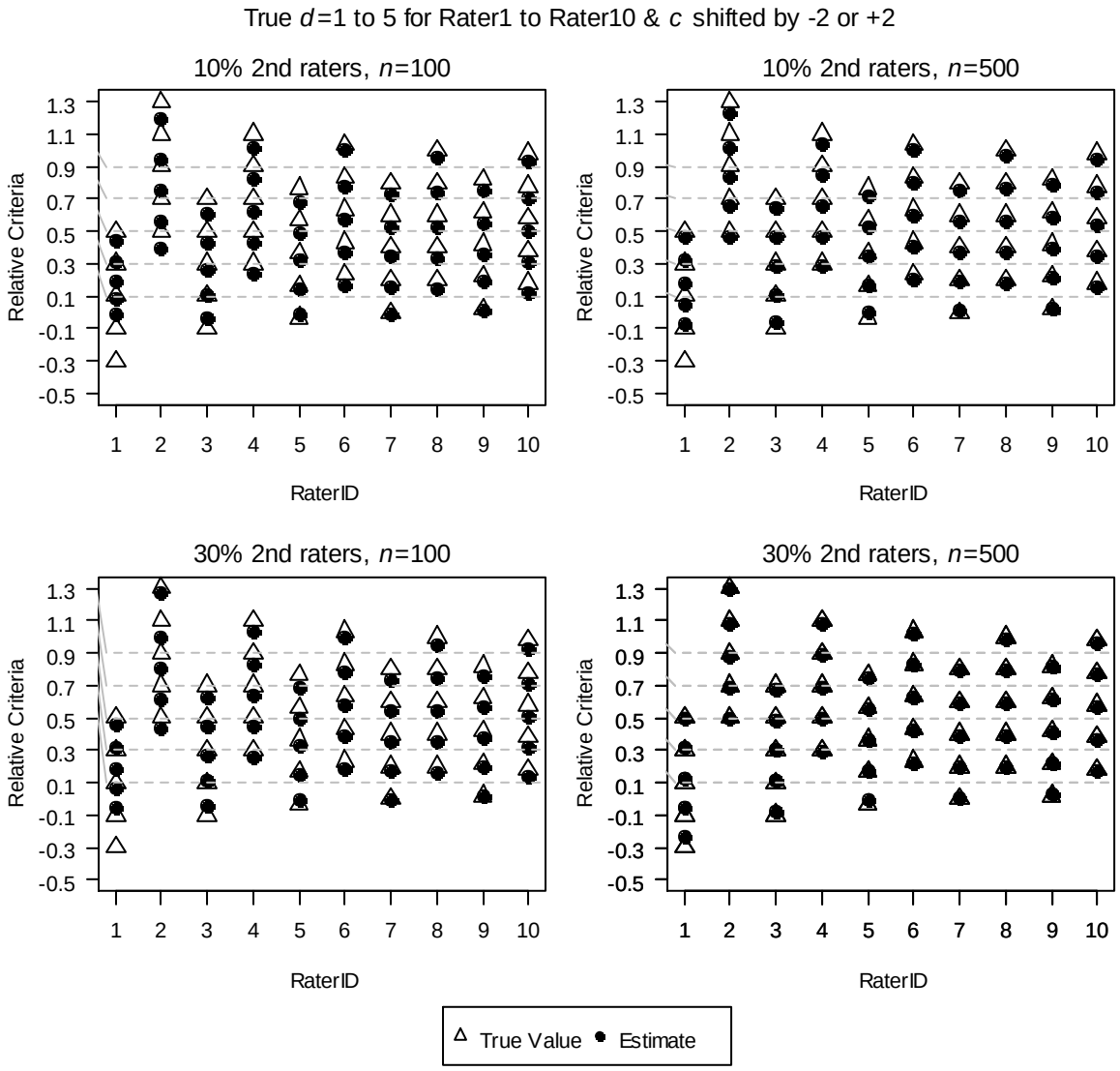


Note. Numbers (1 to 10) indicate the rater ID's

The magnitudes of bias in the detection parameters in the condition with the largest sample size (i.e., 30% second rater with $\bar{n}_j$=500) are moderate (less than 20%), which is similar to the PME results; however, the bias in the other conditions is substantially large (greater than 20%), especially for raters with population detection values that are less than 3 (Rater 1 to Rater 5). Compared to the PME approach, the ranges of MSE are narrower in the Bayesian approach,

and the largest MSE values are found for the smallest detection value with a negative criteria shift, that is, Rater 1.

Figure IV.9. Relative Criteria for Bayesian Estimation with Normal Priors, Simulation 2

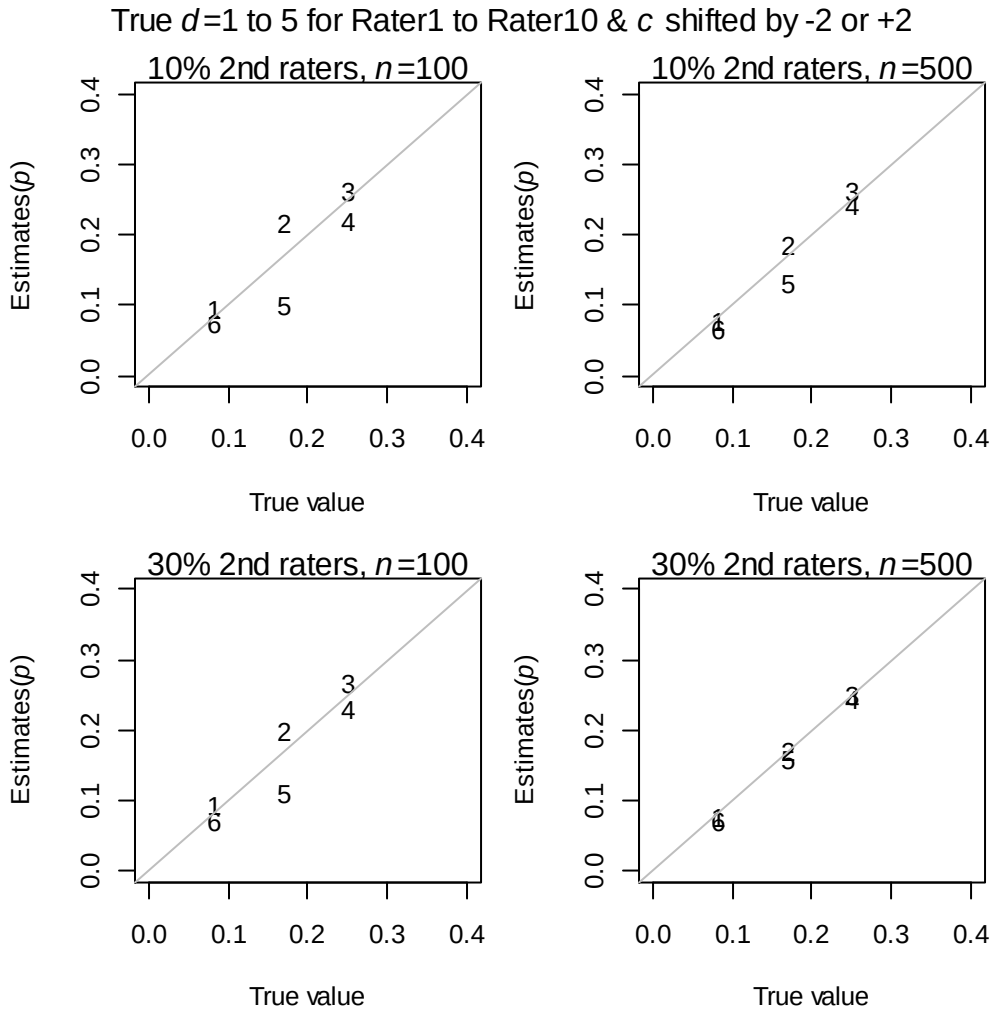


Note. The dotted lines at the mid-point locations indicate no shift in rater criteria.

Similar to the results for the detection parameters, the rater criteria estimates show shrinkage toward to the prior—the estimates are positively biased when the true detections are low (e.g., *d*=1 to 3), but negatively biased when the true detections are high (e.g., *d*=4 to 5). The relative criteria estimates, shown in Figure IV.9, show good recovery of the true values in the

condition with the largest sample size (i.e., 30% second rater with $\bar{n}_j$=500). In the conditions with smaller sample sizes, the relative criteria are negatively biased (i.e., under-estimation) but still reveal the criteria shifts in most cases; which is consistent with the results from the simulation *without* back-readings in the varied detection condition. Similar to the results from the simulation *without* back-readings, the outcome *with* back-readings show shrinkage towards the middle for the relative criteria for raters with low detection and negative criteria shifts (i.e, Rater 1 and Rater 3) when the sample sizes are small (i.e., less than the largest sample size condition).

Figure IV.10. Latent Class Sizes for Bayesian Estimation with Normal Priors in Simulation 2



Note. Numbers (1 to 5) indicates the latent classes

With respect to latent class sizes, shown above in Figure IV.10, the estimates show good recovery (less than 20% bias) when the average number of constructed responses per rater is 500; whereas the estimates of Class 1, 2, and 5 show substantial bias when $\bar{n}_j$ =100. The magnitudes of the bias are within a range of −0.07 to 0.06 (see Appendix B for details), which are similar to those found in the simulations *without* back-readings for Bayesian estimation (i.e., Simulation 1).

Standard Errors/Posterior Standard Deviations

PME. Appendix C2 presents tables that show estimates of the standard errors for the latent class sizes and the rater detection parameters for the partial second rater designs for PME. The impact of the sample sizes on estimates of the standard errors is noticeable.

For the latent class sizes, when $\bar{n}_j$ =500, the bias in SE is generally small to moderate (less than 20%) for all of the classes, which indicates that the standard errors are reasonably well estimated. However, when $\bar{n}_j$ =100, all of classes show large bias (greater than 20%) when there are only 10% second raters. In the condition with 30% second raters and $\bar{n}_j$ =100, only the first class shows a substantial bias.

With regards to the rater detection parameters, when $\bar{n}_j$ =100, most of the biases are large (greater than 20%) and positive, which indicates over-estimation of the SEs. When $\bar{n}_j$ =500, the biases for SE's for the rater detection parameters are large and positive only if the population rater detection values are larger than 3 (i.e., Rater 5 thru Rater 10)—which is similar to the results found in a previous simulation study with an unbalanced design and two raters (DeCarlo, 2010).

Bayesian estimation with informative priors. Appendix D2 presents tables that examine the estimates of the standard deviations for the posterior distributions for the latent class sizes and for the detection parameters for the partial second rater design.

With regards to the latent class sizes, the bias in the SDs is generally small to moderate (less than 20%) except for the last class (i.e., Class 6), where the average number of constructed responses per rater ($\bar{n}_j$) is 500. However, when $\bar{n}_j$ =100, all of classes show large bias (greater than 20%) regardless of the proportion of second raters.

For the rater detection parameters, when $\bar{n}_j$ =100, most of the biases are large (greater than 20%) and positive, which indicates that SDs are over-estimated. When $\bar{n}_j$ =500, the bias for the SDs are large and positive for population detection values that are larger than 3 (i.e., Rater 5 thru Rater 10). These patterns are very similar to the results found above for PME.

Classification

Classification accuracy (proportion correctly classified) for the four conditions are shown in Table IV.2. In addition to the statistics shown in the previous section, the current section includes PC_{av}, which indicates the proportion of cases that were correctly classified in the simulation by using the obtained average score (rounded both up and down; the rounding that gave the largest value of PC_{av} is the one that is reported); the Pearson correlation r_{av} and tau-b_{av} are also presented, which reflect the association between the true latent classes and the average scores.

As shown in Table IV.2, the estimated proportion correctly classified, PC_{pred}, tends to overestimate the proportion actually correctly classified in the simulation (i.e., PC_{obt}). The magnitude of overestimation is moderate in the largest sample size condition (0.03), but it is large in the smallest sample size condition (0.3). Overestimation of PC_{obt} by PC_{pred} has been reported in similar simulation studies (DeCarlo, 2008, 2010).

The proportion correctly classified using the average score, PC_{av}, is lower than the proportion correctly classified using the model, PC_{obt} (around 10%) only in the largest sample

size condition (i.e., 30% second rater with n=500). On the other hand, the correlation statistics (i.e., Pearson *r* and Kendall's τ) using the model are higher than those for the average scores except in the condition with the smallest sample size (i.e., 10% second rater with n=100).

Table IV. 2.

Estimated and Obtained Proportion Correct and Correlations with True Latent Classes in a Partial Second Rater Design (Simulation 2)

Number of Examinees ($\bar{n}_j$)	PC_{pred}	PC_{obt}	PC_{av}	λ_{pred}	λ_{obt}	r_{obt}	r_{av}	τ_{b_obt}	τ_{b_av}
10% second raters									
100	0.603	0.334	0.439	0.415	0.101	0.615	0.755	0.562	0.657
500	0.566	0.445	0.439	0.408	0.256	0.769	0.754	0.700	0.657
30% second raters									
100	0.598	0.458	0.456	0.447	0.269	0.815	0.767	0.746	0.671
500	0.573	0.544	0.457	0.427	0.388	0.848	0.767	0.773	0.671

Note. PC_{pred} = estimated proportion correct; PC_{obt} = obtained (in the simulation) proportion correct; PC_{av} = obtained (in the simulation) proportion correct using the average score; λ_{pred} and λ_{obt} are the estimated and the obtained lambda; r_{obt} and $\tau_{b\text{-}obt}$ are the obtained Pearson correlation and tau-b; r_{av} and $\tau_{b\text{-}av}$ are the obtained Pearson correlation and tau-b for the average scores.

Summary of Simulations

Using parameter constraints (that is specifying the latent class sizes and the rater detection parameters, and estimating only a rater criteria parameter) for a single rater design in Simulation 1 (with only one rater for each constructed response) led to moderate parameter recovery and classification accuracy. In particular, the criteria shifts were recovered by the criteria estimates (except when rater detection values were low).

In Simulation 1, the Bayesian approach also led to recovery of the criteria shifts and gave estimates of the latent class sizes that were close to the true values. However, the detection

parameters were not recovered very well in any condition. Also, the posterior SDs for almost all of the parameters were over-estimated.

The second set of simulations showed that it is useful to use back-reading observations, in that they make the model identified. The results for back-readings in Simulation 2 are moderate for both PME and Bayesian methods, except for the condition with the smallest sample size (i.e., 10% second rater with $\bar{n}_j$=100). In that case, the results showed similar patterns to the results for Simulation 1 (without back-readings). However, PME (without the parameter constraints) performed worse for Simulation 2 as compared to Simulation 1 (PME with the parameter constraints). Specifically, the bias for the rater criteria estimates was substantially larger and the shifts in the criteria parameters (indicating rater effects) were not recovered for the smallest sample size (i.e., 10% second rater with $\bar{n}_j$=100) in Simulation 2, whereas most of rater criteria parameters were recovered fairly well in Simulation 1.

For conditions with larger sample sizes (e.g., more second raters or more constructed responses), there is moderate recovery of most of the rater shifts and the latent class sizes (especially for the condition with the largest sample size—30% second rater with $\bar{n}_j$=500). In Simulation 2, the parameter estimates for Bayesian estimation were closer to the true values than those obtained with PME. Both approaches, however, recovered the parameters very well for the largest sample size.

The impact of sample size was also noticeable with respect to estimating standard errors or posterior standard deviations. With regards to the latent class sizes, when $\bar{n}_j$ = 500, the bias in the estimates of the SEs or SDs was generally small to moderate (less than 20%) except for the last class (i.e., Class 6). However, when $\bar{n}_j$=100, large bias (greater than 20%) was found in all cases.

In addition, overestimation of the proportion correctly classified (PC) was found. The magnitude of the overestimation was moderate for the largest sample size, but was large for the smallest sample size.

IV.3. Results for the Empirical Study

PIRLS 2006 reliability data is used to examine PME and the Bayesian approach for a partial second rater design. Ratings for 7 to 8 raters on two constructed response items with four response categories are analyzed (see Table III.7 for details). For Item 4, with the highest agreement level (e.g., tau=0.93), the responses from 1023 examinees were reviewed by 8 raters and 23% of their responses were reviewed by a second rater (one rater, Rater 9, does not have any linkage to the other raters). For Item 5, with the lowest agreement level (e.g., tau=0.93), 991 examinees' responses were reviewed by 7 raters (i.e., no observation from Rater 9) and 20% of them were used to link the raters.

Item 4 was "give three ways penguins are able to keep warm in Antarctica" accompanied by reading passages regarding Antarctica. The scoring rubric was 3 for 'Extensive Comprehension', 2 for 'Satisfactory Comprehension', and 1 for 'Minimal Comprehension'. A score of 0 indicated no comprehension or no response. Item 5 had to do with "what Da Vinci learned", however, the actual item question is not available to the public; the scoring rubric was the same.

The analysis compares the use of PME and Bayesian estimation with informative priors. Specifically, for PME, Bayes constants of 1 were used. For Bayesian estimation, informative normal priors with means of 3 for d and 1.5 for c_1 and variances of 4 were used. The magnitudes of MC errors for Bayesian estimation were less than 5% of their corresponding SDs, which indicates a sufficient number of MCMC samples. Trace plots for the model parameters in the

application also showed good convergence after 5000 burn-ins and 30,000 updates (as shown in Appendix F). The results for the two approaches are compared. Sensitivity of the results to the priors is also examined.

Parameter Estimation

The parameter estimates and their SE's or SD's are shown in Table IV.3 and Table IV.4 for Item 4 and Item 5, respectively. The latent class size estimates for Item 4 show that Class 4 has the largest size and that the class sizes decrease as one goes from Class 4 to 1. In other words, the latent class size estimates for Item 4 for the LC-SDT model show that the classes are basically negatively skewed (e.g., 50% of responses belong to the highest class). For Item 5, Class 2 has the largest estimated size. These results suggest that Item 4 was an easier item, in that most of the examinees ended up in the highest class. This result is consistent with results found for an earlier analysis of this data using the generalized partial credit model (Mullis, et al., 2006).

With regards to rater parameters, detection varies across the raters and the estimates are uniformly high (i.e., larger than 3). For example, for Item 4, the estimates of d_j range from 2.9 to 7.4 (with SEs of 1.1 to 1.5) for PME and 4.5 to 6.2 (with SDs of 0.8 to 1.0) for Bayesian estimation. Note that Rater 9 does not have any observations to link to the other raters and so detection cannot be estimated with PME for this rater (the output simply gives a value of 0 for this rater), whereas the Bayesian approach provides an estimate that is close to the prior mean of 3. The SEs for the detection parameters are relatively large, whereas the SDs for the Bayesian approach are smaller.

Table IV.3

Parameter Estimates for Item 4

	PME		MCMC		
	Estimate	SE	Estimate	Posterior SD	MC error
Latent Class Size					
Class 1	0.098	0.012	0.096	0.011	0.000
Class 2	0.146	0.018	0.152	0.017	0.000
Class 3	0.253	0.022	0.255	0.021	0.000
Class 4	0.504	0.020	0.497	0.020	0.000
Rater Detection (*d*)					
Rater 1	7.442	1.503	6.228	0.802	0.019
Rater 2	6.524	1.424	5.708	0.763	0.025
Rater 3	6.891	1.439	5.885	0.779	0.026
Rater 4	2.860	1.304	4.540	1.033	0.026
Rater 5	5.744	1.071	5.654	0.758	0.023
Rater 6	6.295	1.327	5.795	0.791	0.028
Rater 7	6.294	1.358	5.803	0.821	0.028
Rater 9	-0.001	1.408	3.273	1.110	0.022

Table IV.4

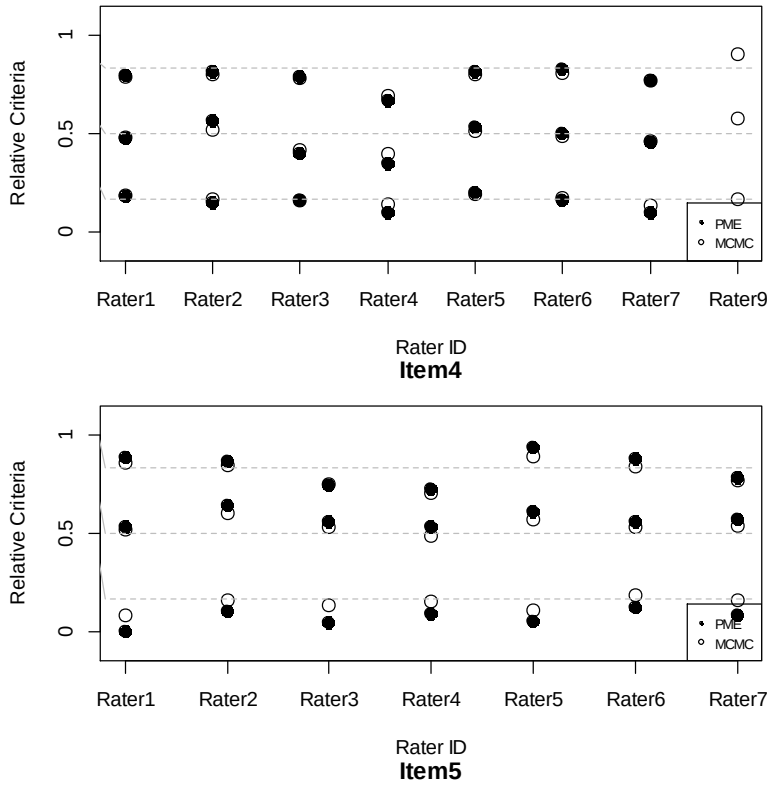
Parameter Estimates for Item 5

	PME		MCMC		
	Estimate	SE	Estimate	Posterior SD	MC error
Latent Class Size					
Class 1	0.196	0.029	0.141	0.022	0.001
Class 2	0.566	0.033	0.639	0.027	0.001
Class 3	0.162	0.022	0.147	0.019	0.000
Class 4	0.076	0.014	0.074	0.013	0.000
Rater Detection (*d*)					
Rater 1	5.215	1.565	4.676	0.891	0.026
Rater 2	5.757	1.547	4.965	0.791	0.027
Rater 3	6.180	1.616	5.695	0.850	0.033
Rater 4	3.673	1.680	4.844	1.075	0.027
Rater 5	5.466	1.493	4.807	0.836	0.026
Rater 6	5.999	1.408	5.577	0.873	0.036
Rater 7	4.885	1.333	5.287	0.955	0.038

The relative criteria locations for Item 4 and Item 5 are shown in Figure IV.11. For Item 4, the raters' criteria are generally quite close to the 'no bias' locations (dotted lines). However, the figure shows that Rater 4 is lenient compared to the other raters, in that the raters criteria all fall below the dotted lines; this is also true, though to a lesser extent, for Rater 7.

For Item 5, the relative criteria estimates from the two methods (PME or Bayesian) differ slightly more than for Item 4, but are still consistent with each other. The criteria estimates also show other types of rater effects that do not appear for Item 4. For example, for Rater 5, the bottom circle is below the dotted line whereas the top circle is above the dotted line. This shows 'central tendency', in that the rater tends to avoid assigning a score of 0 or a score of 3 on the four point scale. This type of rater effect has also been found for other large scale assessments (DeCarlo et al., 2011).

Figure IV.11. Relative Criteria Estimates (*c*)



In addition, computational times for the two methods are very different. While latentGOLD for PME only take a second, Openbugs for MCMC takes 40 mins (Item 5) to 45 mins (Item4).

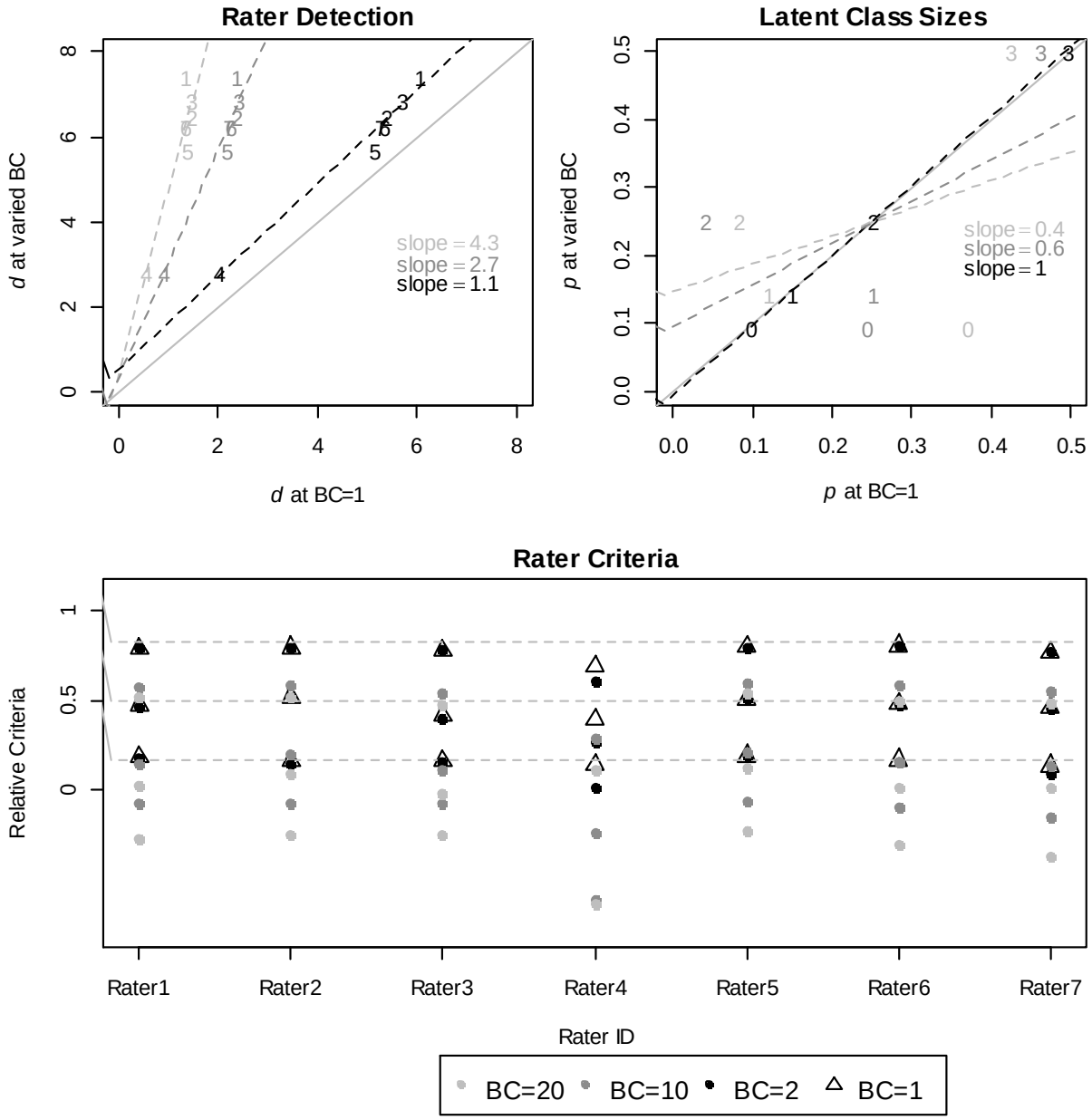
Sensitivity to Priors

A sensitivity analysis to the priors was also conducted. First, effects of the magnitudes of the Bayes' constants in PME were examined. Then, effects of varying the variance of the normal priors in the Bayesian approach were examined.

Sensitivity to Bayes' constants in PME. Bayes constants of 1, 2, 10, and 20 were used. Figure IV.12 shows the results for the rater parameters and relative criteria locations, with results for a Bayes constant of 1 used as the reference.

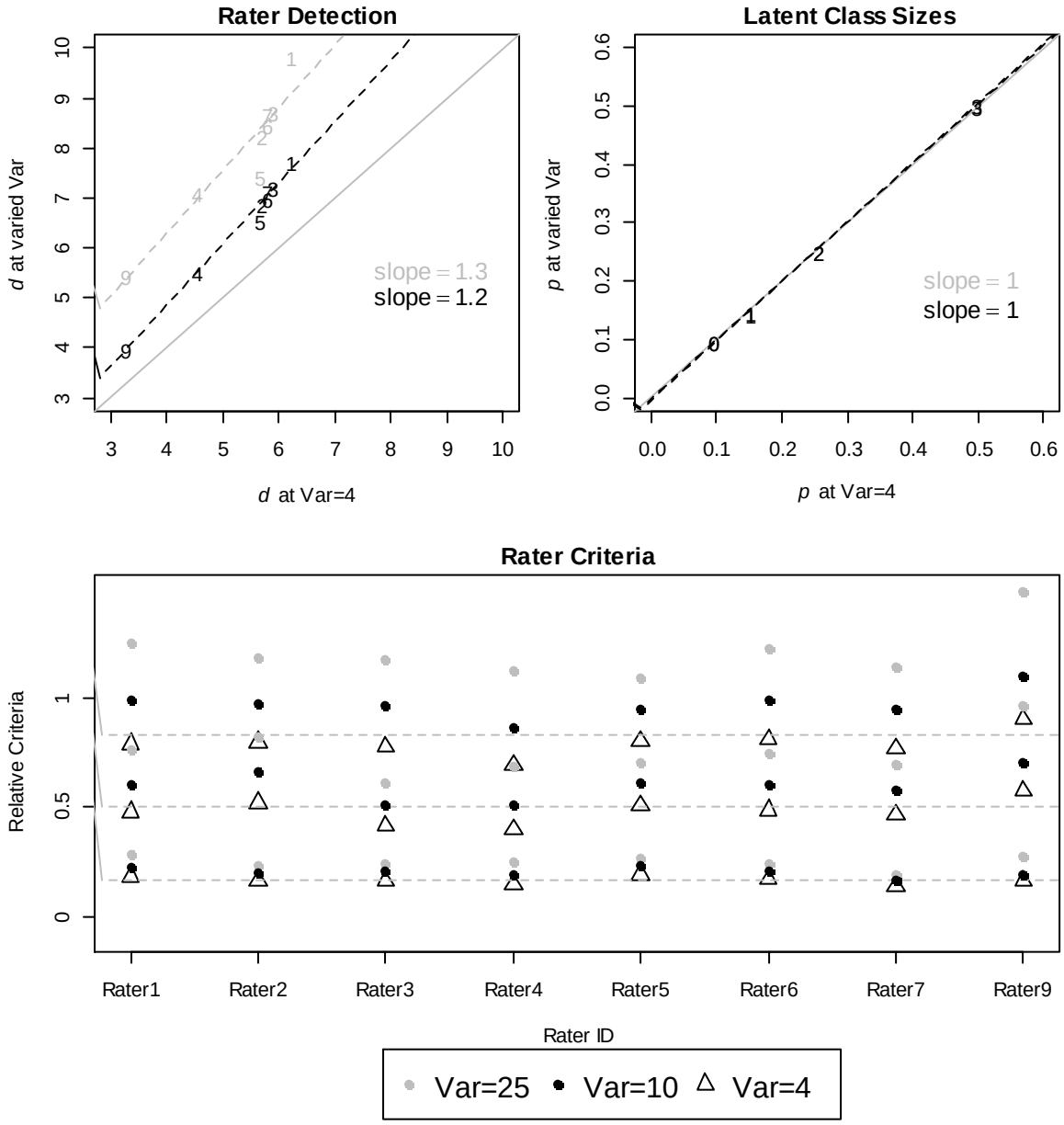
The results for the detection parameter estimates show the expected shrinkage towards zero as the Bayes constant increases. Similarly, it is apparent that the latent class sizes are smoothed towards the uniform distribution, in that the class size estimates for the lowest category increase as the Bayes constants increase and those for the highest category decrease. The relative criteria estimates for the lowest response category are also sensitive to the magnitude of the Bayes constant, whereas the other response categories appear to be less effected. One exception to these patterns is Rater 4, who scored the smallest number of examinees (48).

Figure IV.12. Effects of Bayes' constants in PME



Sensitivity to normal prior variance, Bayesian estimation. Variances of 4, 10, and 25 were used, and the results are shown in Figure IV.13. These values were selected based on a previous study with small sample sizes in the context of logistic regression (Galindo-Garre et al., 2004). Mean values of 3 for d and 1.5 for c_{j1} were used, as above, and the variance of 4 (which is the one used earlier) was used as the reference point.

Figure IV.13. Effects of Normal Prior Variances for Bayesian Estimation



For detection, as the variance increased, the mean detection estimates increased. For the latent class sizes, there is no effect of changing the variance of the prior. For the relative criteria, the locations for the lowest category were not sensitive to the variance, whereas the locations of the highest category were sensitive and were shifted upwards as the variance got larger.

Summary of the Empirical Study

The analysis demonstrates the utility and the application of LC-SDT for the PIRLS USA reliability data. The analysis provides useful information about the latent class sizes, rater discrimination, and various rater effects, as reflected by the response criteria locations. For example, the analysis clearly revealed rater effects that were consistent across PME and Bayesian estimation.

The conclusions were also basically unchanged when the priors were varied. However, the effects of varying the Bayes constants in PME and the variance of the normal priors in the Bayesian approach were somewhat different. This is expected because of differences between the approaches. For example, the detection values are shrunk towards zero in PME whereas they are shrunk towards the priors in Bayesian estimation. Further study of the effects of Bayes constants in PME and different priors in Bayesian estimation in the context of LC-SDT is needed.

Chapter V

SUMMARY AND DISCUSSION

V.1. Summary and Discussion

The present study examined the performance of LC-SDT models in sparse rater designs. Particularly, the paper explored the utility of using parameter constraints, Bayesian estimation, and back-readings in situations where model identification issues arise. The performances of the approaches were examined with simulations and with an empirical study.

With respect to simulations, two main simulations were conducted. The first simulation examined the performance of the LC-SDT model when it was used with only one rater assigned to each constructed response. Models with different sample sizes (e.g., number of constructed responses per rater) and different rater effects (e.g., severity or leniency) were examined.

The results showed that using parameter constraints (specifying the latent class sizes and rater detections, and only estimating the rater criteria locations) gave moderate parameter recovery. Although the bias for both PME and Bayesian estimation was relatively large, both approaches were able to detect whether a rater was strict or lenient, which is of major concern in real-world research. The Bayesian approach offers the advantage of being able to uncover differences in latent class sizes and rater effects.

Classification accuracy was also reasonable, though it was relatively low (about 50%) when only one rater was used for each constructed response. The important implication is that more than one rater is really needed to get classification accuracy up to an acceptable level, except possibly in situations where rater detection is very high (as found for some constructed responses, such as for mathematics items; see DeCarlo, 2010).

The second simulation examined parameter recovery when back-readings are available. The simulation also examined the effects of having back-readings available for a different percentage of cases. Back-reading observations are collected by many testing agencies to estimate inter-rater agreement or reliability, but they have not been used for anything else. It is shown here that the back-readings can be very useful, in that they allow one to fit the LC-SDT model. In that case, one obtains information not only about rater reliability (via the rater discrimination parameter) but also about rater effects and latent class sizes. Information about rater effects is particularly relevant to issues concerning rater training, monitoring, and selection. The results showed that the rank order of rater detection was close to the true order and the relative criteria estimates successfully revealed rater effects (e.g., leniency or severity). The simulation results also showed that increasing the percentage of back-readings from 10% to 30% led to a considerable improvement in estimation. Thus, testing companies should consider increasing the percentage of back-readings, or at least determine what the largest cost-effective percentage could be.

The Bayesian approach offers the advantage of allowing one to incorporate previous knowledge or beliefs. Although informative priors with a variance of 4 were used in the current paper, using even tighter variance can be justified. For example, DeCarlo (2010) found, for an analysis of a real-world large scale assessment, that the variance of the detection parameters across raters was less than 1. Using a smaller prior variance might improve estimation in conditions where estimation was marginal, such as for the smallest sample sizes. This should be examined in future research.

V.2. Limitations and Future Research

The present study has several limitations that could be addressed in future research. For example, the current study investigated the utility of the parameter constraint method where the true values of latent class sizes are known and correctly specified. In order to illustrate the full utility of the approach, one should examine the sensitivity to latent class size misspecification. Also, the current study only examined situations where the latent class sizes were discretely normally distributed, however it would be beneficial to use other distributions in future studies. One of the merits of LC-SDT includes its' non-parametric capacity to describe latent class distribution (as shown in the empirical study). For example, the utility of the proposed methods can be examined in a negatively skewed distribution (which are often observed in many large scale tests; DeCarlo, 2008).

Results of the present study are also specific to the population values used in the simulation. While the current study included rater detection values of 1 to 5, which have been found in many applications, some applications (with high inter-rater agreement levels), such as the PIRLS study examined here, have even higher levels of rater detection. Thus, it would be useful to examine the approaches with different values of detection. Specially, there have been reported applications with higher detection values in large scale tests. For example, DeCarlo (2010) found that in a mathematics test, rater detection values ranged from 8 to 12.

For the approach with parameter constraints, one could argue that the approach is limited because one really doesn't need a model in that case and can estimate rater effects just using the observed scores. However, an important advantage of the model based approach is that the simple model can be extended in various useful ways, such as to include covariates (Wang,

2012), multiple-choice items (Kim, 2009), or multiple CR items (DeCarlo et al, 2011). Thus, the approach might be useful in other contexts and should be further explored.

In addition to the proposed solutions, other methods to resolve the issues in sparse rater situations need to be investigated in future. For instance, raters can be considered as items in equating studies where back-readings can be viewed as anchor items. In this case, methods to equate or link items where no anchor items (i.e., common items) are present would be similarly applied to single rater designs where there are no back-readings. Another approach that needs to be considered includes matrix completion methods (e.g.,Candes & Romberg, 2007) that try to recover a lower-rank matrix from a sampling of its entries.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1,* 201-218.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Agresti, A. , & Yang, M. (1986). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis, 5,* 9–21.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barnes, L. L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4,* 143-157.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of changes. *Philosophical Transactions of the Royal Society, 53,* 370-418.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of inter rater agreement measures. *Canadian Journal of Statistics 27,* 3-23.

Bollen, K. A. (1989), *Structural equations with latent variables*. New York: John Wiley & Sons, Inc.

Boughton, K., Klinger, D., & Gierl, M. (2001). *Effects of random rater error on parameter recovery of the generalized partial credit model and graded response model.* Paper presented at the Annual Meeting of the National Council of Measurement in Education, Seattle, WA.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City IA: ACT.

Candes, E.J. & Romberg, J. (2007). Sparsity and incoherence in compressive sampling. *Inverse Problems, 23*(3), 969-985.

Cao, J., Stokes, S. L., & Zhang, S. (2010). A Bayesian approach to ranking and rater evaluation: An application to grant reviews. *Journal of Educational and Behavioral Statistics, 35,* 194-214.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311-359). New York: Plenum Press.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., & Wiedman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association, 86,* 68–78.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Curtis, M. S. (2010). BUGS code for item response theory. *Journal of Statistical Software. 36*(1), 1-34.

Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage Publications.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73,* 533-559.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory with applications. *Multivariate Behavioral Research, 37,* 423-451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42,* 53-76.

DeCarlo, L. T. (2008). *Studies of a latent-class signal-detection model for constructed response scoring* (ETS Research Rep. No. RR-08-63). Princeton NJ: ETS.

DeCarlo, L. T. (2010). *Studies of a latent-class signal-detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report No. RR-10-08). Princeton NJ: ETS.

DeCarlo, L. T., & Kim, Y.K. (2008, March). *Score resolution in essay grading: A view from a signal detection model of rater behavior.* Paper presented at the 2008 annual meeting of the American Educational Research Association, New York, NY.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement, 48,* 333-356.

De Leeuw, J., van der Heijden, P.G.M., & Verboon, P. (1990). A latent time budget model. *Statistica Neerlandica, 44*, 1-22.

Donoghue, J. R., & Hombo, C. M. (2000). *A comparison of different model assumptions about rater effects.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. April 2000, New Orleans, LA.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Engelhard, G. Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). New York: The College Board.

Fleiss, J. L., Levin, B., & Paik, M.C. (2003). *Statistical methods for rates and proportions (3rd ed.)*. New York: John Wiley & Sons.

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. NY: Springer.

Galindo-Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika, 33,* 43-59

Galindo-Garre, F., Vermunt, J. K., & Bergsma, W.P. (2004). Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods & Research, 33*(1), 88-117.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis.* London: Chapman & Hall.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215-231.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications, Part I. *Journal of the American Statistical Association, 49*, 732–764.

Greenland, S., Schwartzbaum, J. A., & Finkle,W. D. (2000). Problems from small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology, 151*, 531–539.

Hayduk. L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore: Johns Hopkins Press.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (ETS Research Rep. No. RR-01-05). Princeton, NJ: ETS.

Hui, S. L., & Zhou, X. H. (1998). Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research, 7*, 354-370.

Jackman, S. (2004). What do we learn from graduate admissions committees?: A multiple-rater, latent variable model, with incomplete discrete and continuous indicators. *Political Analysis, 12*(4): 400-424.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester: Wiley.

Jones, M. & Vickers, D. (2011). Considerations for performance scoring when designing and developing next generation assessments. *Pearson's White Papers*. Retrieved from http://www.pearsonassessments.com/hai/images/tmrs/Performance_Scoring_for_Next_Gen_Assessments.pdf

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998), Markov chain Monte Carlo in practice: A roundtable discussion. *Statistical Science, 52*(2), 93–100.

Kenny, D. A., Kashy, D., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, and G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., pp. 233-265). New York: McGraw-Hill.

Koehler, E. Brown, E. & Haneuse, S. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician, 63*(2), 155-162.

Kim, Y. K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model.* Ph.D. dissertation, Teachers College, Columbia University, NY.

Lee, S.-Y. & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653–686.

Lee, M. D., & Wagenmakers, E.-J. (2010). A course in Bayesian graphical modeling for cognitive science. Course notes, University of California Irvine. Retrieved from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf

Levy, R. (2006). *Posterior Predictive Model Checking for Multidimensionality in Item Response Theory and Bayesian Networks*. Ph.D. dissertation, University of Maryland.

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics, 537139,* 1-18.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: theory into practice* (Vol. 3, pp. 85-112). Norwood, NJ: Alex Publishing Corporation.

Lord, F. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48,* 425-435.

Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments.* Ph.D. dissertation, Carnegie Mellon University, PA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McHugh, R. (1958). Note on: Efficient estimation and local identification in latent classes analysis. *Psychometrika, 23*, 273-274.

McLachlan, G. J. & Peel, D. (2000). *Finite mixture models.* NY: Wiley.

Mullis, I, Martin, M., Kennedy, A. & Foy, P. (2006). *PIRLS 2006 international report*. Boston: International Study Center, Boston University.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Murphy, K. R., & Balzer, W K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619-624.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

Nelson, J. C. & Pepe, M. S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research, 9*(5), 475-96.

Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory and item response theory.* Ph.D. dissertation, Teachers College, Columbia University, NY.

Parshall, C. G., Kromrey, J. D., & Chason, W. M. (1996). *Comparison of alternative models for item parameter estimation with small samples.* Paper presented at the Annual Meeting of the Psychometric Society, Banff, Alberta, Canada.

Patz, R. J. & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342–366.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

Plummer, M., Best, N. G., Cowles, K., & Vines, K. (2011). *Coda: Output analysis and diagnostics for MCMC.* R package version 0.14-6. Retrieved from http://cran.r-project.org/web/packages/coda/index.html

Qu, Y. , Tan, M., & Kutner, M. H. (1996). Random effects models in latent class analysis forevaluating accuracy of diagnostic tests. *Biometrics, 52,* 797–810.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark Paedagogiske Institute (Reprinted by University of Chicago Press, 1980).

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-92.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Samejima, F. (1969). Estimation of latent ability using a response of graded scores. *Psychometrika Monograph No*. *17*, 34.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall.

Setiadi, H. (1997). *Small sample IRT item parameter estimates*. Ed.D. dissertation, University of Massachusetts Amherst, United States, Massachusetts.

Shaw, S. (2004). IELTS Writing: revising assessment criteria and scales (Phase 3), *Research Notes, 16,* 3–7.

Smith, M. K. & Richardson, H. (2007). WinBUGSio: A SAS macro for the remote execution of WinBUGS. *Journal of Statistical Software, 23*(9), 1-10.

Spiegelhalter, D., Thomas, A., Best, N. & Gilks, W. (1996). *BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual (version ii).* MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2011). *OPENBUGS User Manual (Version 3.2.1).* Cambridge, UK: MRC Biostatistics Unit.

Sykes, R. C., Ito, K. & Wang, Z. (2008). Effects of assigning raters to items. *Educational Measurement: Issues and Practice, 27,* 47–55.

Tanner, M. A. & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Psychological Association, 80,* 175-180.

Taylor, L. & Jones, N. (2001). Revising the IELTS speaking test, *Research Notes, 4,* 9–12.

Uebersax, J. S. (2012a). *What is model identifiability and how important an issue is it?* Retrieved from http://www.john-uebersax.com/stat/faq.htm#ident

Uebersax, J. S. (2012b). *The myth of chance-corrected agreement.* Retrieved from http://www.john-uebersax.com/stat/kappa2.htm

Uebersax, J. S. & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine, 9,* 559-572.

van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*(4), 519–538.

Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. Van Duijn, and T. A. B. Snijders (Eds.), *Essays on item response modeling* (pp. 89-108). New York: Springer-Verlag.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent Gold 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2007). *LG-Syntax™ user's guide: Manual for Latent Gold 4.5 syntax module.* Belmont, MA: Statistical Innovations Inc.

von Eye, A., & Mun, E.Y. (2005). *Modeling rater agreement - manifest variable approaches*. Mahwah, NJ: Lawrence Erlbaum.

Walter, S. D. & Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology, 41*, 923-937.

Wang, Z.G. (2012). *On the use of covariates in a latent class signal detection model, with applications to constructed response scoring.* Ph.D. dissertation, Teachers College, Columbia University, NY.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–287.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wolfe, E. W., Myford, C. M., Engelhard, G. E., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays.* (College Board Research Report No. 2007-2). New York: The College Board.

Appendix A1

Parameter Estimates, Bias, Percent Bias, and MSE for Parameter Constraints (via PME) in Simulation 1 (Single Rater per Examinee)

Table A1.1
Constant Detection (d=2), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	Values	Estimates	bias	%bias	MSE
100	c_{11}	-1.5	-0.619	0.881	58.730	1.423
100	c_{12}	1.5	1.997	0.497	33.130	0.729
100	c_{13}	4.5	4.834	0.334	7.420	0.555
100	c_{14}	7.5	7.566	0.066	0.880	0.421
100	c_{15}	10.5	10.269	-0.231	2.200	0.452
30	c_{21}	-1.5	-0.152	1.348	89.870	2.181
30	c_{22}	1.5	1.745	0.245	16.330	1.613
30	c_{23}	4.5	4.644	0.144	3.200	1.512
30	c_{24}	7.5	7.328	-0.172	2.290	1.067
30	c_{25}	10.5	10.307	-0.193	1.840	1.464
100	c_{31}	0.0	0.380	0.380	.	0.843
100	c_{32}	3.0	3.272	0.272	9.070	0.520
100	c_{33}	6.0	5.938	-0.062	1.030	0.316
100	c_{34}	9.0	8.806	-0.194	2.160	0.417
100	c_{35}	12.0	11.696	-0.304	2.530	0.563
150	c_{41}	0.0	0.368	0.368	.	0.416
150	c_{42}	3.0	3.185	0.185	6.170	0.286
150	c_{43}	6.0	6.138	0.138	2.300	0.231
150	c_{44}	9.0	8.981	-0.019	0.210	0.258
150	c_{45}	12.0	11.744	-0.256	2.130	0.338
50	c_{51}	1.5	1.819	0.319	21.270	1.298
50	c_{52}	4.5	4.721	0.221	4.910	0.863
50	c_{53}	7.5	7.475	-0.025	0.330	0.664
50	c_{54}	10.5	10.304	-0.196	1.870	0.884
50	c_{55}	13.5	13.111	-0.389	2.880	0.939
50	c_{61}	1.5	1.888	0.388	25.870	1.031
50	c_{62}	4.5	4.682	0.182	4.040	0.635
50	c_{63}	7.5	7.579	0.079	1.050	0.546
50	c_{64}	10.5	10.336	-0.164	1.560	0.629
50	c_{65}	13.5	13.347	-0.153	1.130	1.105
170	c_{71}	3.0	3.264	0.264	8.800	0.291
170	c_{72}	6.0	6.098	0.098	1.630	0.176
170	c_{73}	9.0	8.882	-0.118	1.310	0.173
170	c_{74}	12.0	11.689	-0.311	2.590	0.298

Table A1.1 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	Values	Estimates	bias	%bias	MSE
170	c_{75}	15.0	14.472	-0.528	3.520	0.550
150	c_{81}	3.0	3.285	0.285	9.500	0.351
150	c_{82}	6.0	6.037	0.037	0.620	0.245
150	c_{83}	9.0	8.820	-0.180	2.000	0.303
150	c_{84}	12.0	11.661	-0.339	2.830	0.407
150	c_{85}	15.0	14.480	-0.520	3.470	0.875
100	c_{91}	4.5	4.608	0.108	2.400	0.490
100	c_{92}	7.5	7.506	0.006	0.080	0.461
100	c_{93}	10.5	10.369	-0.131	1.250	0.424
100	c_{94}	13.5	13.196	-0.304	2.250	0.614
100	c_{95}	16.5	15.699	-0.801	4.850	1.209
100	c_{101}	4.5	4.716	0.216	4.800	0.445
100	c_{102}	7.5	7.504	0.004	0.050	0.433
100	c_{103}	10.5	10.335	-0.165	1.570	0.439
100	c_{104}	13.5	13.185	-0.315	2.330	0.608
100	c_{105}	16.5	15.758	-0.742	4.500	1.349

Table A1.2
Constant Detection (d=2), $\bar{n}_j$ =500, PME with Parameter Constraints

sample size	Parameter	Values	Estimates	bias	%bias	MSE
500	c_{11}	-1.5	-0.771	0.729	48.600	0.647
500	c_{12}	1.5	1.880	0.380	25.330	0.252
500	c_{13}	4.5	4.698	0.198	4.400	0.129
500	c_{14}	7.5	7.490	-0.010	0.130	0.081
500	c_{15}	10.5	10.330	-0.170	1.620	0.106
150	c_{21}	-1.5	-0.815	0.685	45.670	0.875
150	c_{22}	1.5	1.865	0.365	24.330	0.489
150	c_{23}	4.5	4.748	0.248	5.510	0.359
150	c_{24}	7.5	7.544	0.044	0.590	0.243
150	c_{25}	10.5	10.324	-0.176	1.680	0.358
500	c_{31}	0.0	0.508	0.508	.	0.386
500	c_{32}	3.0	3.258	0.258	8.600	0.161
500	c_{33}	6.0	6.095	0.095	1.580	0.096
500	c_{34}	9.0	8.900	-0.100	1.110	0.089
500	c_{35}	12.0	11.764	-0.236	1.970	0.156
750	c_{41}	0.0	0.450	0.450	.	0.269
750	c_{42}	3.0	3.249	0.249	8.300	0.119
750	c_{43}	6.0	6.034	0.034	0.570	0.048
750	c_{44}	9.0	8.863	-0.137	1.520	0.057
750	c_{45}	12.0	11.720	-0.280	2.330	0.135
250	c_{51}	1.5	1.855	0.355	23.670	0.370
250	c_{52}	4.5	4.686	0.186	4.130	0.219
250	c_{53}	7.5	7.461	-0.039	0.520	0.159
250	c_{54}	10.5	10.247	-0.253	2.410	0.197
250	c_{55}	13.5	13.125	-0.375	2.780	0.334
250	c_{61}	1.5	1.869	0.369	24.600	0.300
250	c_{62}	4.5	4.695	0.195	4.330	0.180
250	c_{63}	7.5	7.471	-0.029	0.390	0.162
250	c_{64}	10.5	10.301	-0.199	1.900	0.185
250	c_{65}	13.5	13.127	-0.373	2.760	0.339
850	c_{71}	3.0	3.245	0.245	8.170	0.111
850	c_{72}	6.0	6.100	0.100	1.670	0.053
850	c_{73}	9.0	8.928	-0.072	0.800	0.041
850	c_{74}	12.0	11.740	-0.260	2.170	0.121
850	c_{75}	3.0	14.510	-0.490	3.270	0.311
750	c_{81}	6.0	3.238	0.238	7.930	0.108
750	c_{82}	9.0	6.072	0.072	1.200	0.057
750	c_{83}	12.0	8.913	-0.087	0.970	0.047
750	c_{84}	15.0	11.727	-0.273	2.280	0.132

Table A1.2 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =500, PME with Parameter Constraints

sample size	Parameter	Values	Estimates	bias	%bias	MSE
750	c_{85}	15.0	14.472	-0.528	3.520	0.372
500	c_{91}	4.5	4.716	0.216	4.800	0.121
500	c_{92}	7.5	7.489	-0.011	0.150	0.081
500	c_{93}	10.5	10.295	-0.205	1.950	0.122
500	c_{94}	13.5	13.137	-0.363	2.690	0.231
500	c_{95}	16.5	15.713	-0.787	4.770	0.747
500	c_{101}	4.5	4.739	0.239	5.310	0.140
500	c_{102}	7.5	7.523	0.023	0.310	0.089
500	c_{103}	10.5	10.361	-0.139	1.320	0.120
500	c_{104}	13.5	13.099	-0.401	2.970	0.273
500	c_{105}	16.5	15.752	-0.748	4.530	0.699

Table A1.3
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	Values	Estimates	bias	%bias	MSE
100	c_{11}	-4.5	-0.343	4.157	92.380	17.912
100	c_{12}	-1.5	1.312	2.812	187.470	8.522
100	c_{13}	1.5	3.202	1.702	113.470	3.414
100	c_{14}	4.5	5.300	0.800	17.780	1.061
100	c_{15}	7.5	7.542	0.042	0.560	0.326
30	c_{21}	7.5	7.509	0.009	0.120	1.410
30	c_{22}	10.5	9.754	-0.746	7.100	1.859
30	c_{23}	13.5	11.853	-1.647	12.200	4.142
30	c_{24}	16.5	13.554	-2.946	17.850	10.298
30	c_{25}	19.5	14.853	-4.647	23.830	22.298
100	c_{31}	-1.5	-0.751	0.749	49.930	1.065
100	c_{32}	1.5	1.826	0.326	21.730	0.622
100	c_{33}	4.5	4.586	0.086	1.910	0.441
100	c_{34}	7.5	7.432	-0.068	0.910	0.358
100	c_{35}	10.5	10.284	-0.216	2.060	0.383
150	c_{41}	4.5	4.627	0.127	2.820	0.321
150	c_{42}	7.5	7.471	-0.029	0.390	0.301
150	c_{43}	10.5	10.389	-0.111	1.060	0.328
150	c_{44}	13.5	13.135	-0.365	2.700	0.598
150	c_{45}	16.5	15.770	-0.730	4.420	1.061
50	c_{51}	-0.5	-0.450	0.050	10.000	0.633
50	c_{52}	2.5	2.297	-0.203	8.120	1.344
50	c_{53}	5.5	5.443	-0.057	1.040	0.970
50	c_{54}	8.5	8.439	-0.061	0.720	0.776
50	c_{55}	11.5	11.519	0.019	0.170	0.993
50	c_{61}	3.5	3.537	0.037	1.060	0.953
50	c_{62}	6.5	6.613	0.113	1.740	0.802
50	c_{63}	9.5	9.625	0.125	1.320	0.673
50	c_{64}	12.5	12.632	0.132	1.060	1.120
50	c_{65}	15.5	15.445	-0.055	0.350	0.712
170	c_{71}	0.0	-0.313	-0.313	.	0.527
170	c_{72}	3.0	2.827	-0.173	5.770	0.275
170	c_{73}	6.0	5.930	-0.070	1.170	0.185
170	c_{74}	9.0	9.051	0.051	0.570	0.183
170	c_{75}	12.0	12.139	0.139	1.160	0.281
150	c_{81}	3.0	2.862	-0.138	4.600	0.275
150	c_{82}	6.0	5.960	-0.040	0.670	0.226
150	c_{83}	9.0	9.063	0.063	0.700	0.243
150	c_{84}	12.0	12.202	0.202	1.680	0.375
150	c_{85}	15.0	15.230	0.230	1.530	0.512

Table A1.3 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	Values	Estimates	bias	%bias	MSE
100	c_{91}	0.3	0.101	-0.199	66.330	0.630
100	c_{92}	3.3	3.276	-0.024	0.730	0.505
100	c_{93}	6.3	6.246	-0.054	0.860	0.380
100	c_{94}	9.3	9.348	0.048	0.520	0.463
100	c_{95}	12.3	12.399	0.099	0.800	0.496
100	c_{101}	2.7	2.451	-0.249	9.220	0.494
100	c_{102}	5.7	5.570	-0.130	2.280	0.444
100	c_{103}	8.7	8.641	-0.059	0.680	0.445
100	c_{104}	11.7	11.764	0.064	0.550	0.481
100	c_{105}	14.7	14.988	0.288	1.960	0.639

Table A1.4
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	Values	Estimates	bias	%bias	MSE
500	c_{11}	-4.50	-0.499	4.001	88.910	16.107
500	c_{12}	-1.50	1.134	2.634	175.600	7.050
500	c_{13}	1.50	3.110	1.610	107.330	2.685
500	c_{14}	4.50	5.335	0.835	18.560	0.772
500	c_{15}	7.50	7.497	-0.003	0.040	0.087
150	c_{21}	7.50	7.509	0.009	0.120	0.290
150	c_{22}	10.50	9.714	-0.786	7.490	0.827
150	c_{23}	13.50	11.804	-1.696	12.560	3.112
150	c_{24}	16.50	13.735	-2.765	16.760	8.062
150	c_{25}	19.50	15.524	-3.976	20.390	16.329
500	c_{31}	-1.50	-0.732	0.768	51.200	0.704
500	c_{32}	1.50	1.848	0.348	23.200	0.210
500	c_{33}	4.50	4.629	0.129	2.870	0.104
500	c_{34}	7.50	7.459	-0.041	0.550	0.076
500	c_{35}	10.50	10.262	-0.238	2.270	0.132
750	c_{41}	4.50	4.748	0.248	5.510	0.116
750	c_{42}	7.50	7.526	0.026	0.350	0.054
750	c_{43}	10.50	10.344	-0.156	1.490	0.070
750	c_{44}	13.50	13.162	-0.338	2.500	0.179
750	c_{45}	16.50	15.783	-0.717	4.350	0.605
250	c_{51}	-0.50	-0.525	-0.025	5.000	0.287
250	c_{52}	2.50	2.448	-0.052	2.080	0.199
250	c_{53}	5.50	5.467	-0.033	0.600	0.158
250	c_{54}	8.50	8.463	-0.037	0.440	0.142
250	c_{55}	11.50	11.501	0.001	0.010	0.190
250	c_{61}	3.50	3.460	-0.040	1.140	0.126
250	c_{62}	6.50	6.550	0.050	0.770	0.095
250	c_{63}	9.50	9.563	0.063	0.660	0.141
250	c_{64}	12.50	12.617	0.117	0.940	0.176
250	c_{65}	15.50	15.564	0.064	0.410	0.285
850	c_{71}	0.00	-0.263	-0.263	.	0.146
850	c_{72}	3.00	2.872	-0.128	4.270	0.062
850	c_{73}	6.00	5.957	-0.043	0.720	0.043
850	c_{74}	9.00	9.024	0.024	0.270	0.034
850	c_{75}	12.00	12.106	0.106	0.880	0.056
750	c_{81}	3.00	2.839	-0.161	5.370	0.082
750	c_{82}	6.00	5.943	-0.057	0.950	0.058
750	c_{83}	9.00	9.015	0.015	0.170	0.048
750	c_{84}	12.00	12.100	0.100	0.830	0.057
750	c_{85}	15.00	15.238	0.238	1.590	0.111

Table A1.4 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	Values	Estimates	bias	%bias	MSE
500	c_{91}	0.30	0.042	-0.258	86.000	0.179
500	c_{92}	3.30	3.212	-0.088	2.670	0.091
500	c_{93}	6.30	6.332	0.032	0.510	0.080
500	c_{94}	9.30	9.429	0.129	1.390	0.085
500	c_{95}	12.30	12.552	0.252	2.050	0.157
500	c_{101}	2.70	2.429	-0.271	10.040	0.152
500	c_{102}	5.70	5.551	-0.149	2.610	0.103
500	c_{103}	8.70	8.603	-0.097	1.110	0.074
500	c_{104}	11.70	11.782	0.082	0.700	0.070
500	c_{105}	14.70	14.972	0.272	1.850	0.177

Appendix A2

Evaluation of the Estimated Standard Errors for Parameter Constraints (via PME) in Simulation 1 (Single Rater per Examinee)

Table A2.1
Constant Detection (d=2), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
100	c_{11}	0.808	0.817	0.009	1.114
100	c_{12}	0.697	0.707	0.010	1.435
100	c_{13}	0.669	0.638	-0.031	4.634
100	c_{14}	0.649	0.617	-0.032	4.931
100	c_{15}	0.635	0.643	0.009	1.417
30	c_{21}	0.610	1.391	0.782	128.197
30	c_{22}	1.253	1.285	0.033	2.634
30	c_{23}	1.227	1.141	-0.086	7.009
30	c_{24}	1.024	1.087	0.063	6.152
30	c_{25}	1.201	1.147	-0.053	4.413
100	c_{31}	0.840	0.768	-0.072	8.571
100	c_{32}	0.671	0.667	-0.004	0.596
100	c_{33}	0.562	0.599	0.038	6.762
100	c_{34}	0.620	0.600	-0.020	3.226
100	c_{35}	0.689	0.665	-0.024	3.483
150	c_{41}	0.532	0.622	0.089	16.729
150	c_{42}	0.504	0.544	0.040	7.937
150	c_{43}	0.463	0.489	0.026	5.616
150	c_{44}	0.510	0.490	-0.021	4.118
150	c_{45}	0.525	0.544	0.019	3.619
50	c_{51}	1.099	1.008	-0.091	8.280
50	c_{52}	0.907	0.898	-0.009	0.992
50	c_{53}	0.819	0.856	0.038	4.640
50	c_{54}	0.924	0.895	-0.029	3.139
50	c_{55}	0.892	1.000	0.107	11.996
50	c_{61}	0.943	1.002	0.059	6.257
50	c_{62}	0.780	0.898	0.118	15.128
50	c_{63}	0.738	0.862	0.124	16.802
50	c_{64}	0.780	0.902	0.122	15.641
50	c_{65}	1.045	1.016	-0.030	2.871
170	c_{71}	0.473	0.512	0.039	8.245
170	c_{72}	0.410	0.458	0.048	11.707
170	c_{73}	0.401	0.458	0.056	13.965
170	c_{74}	0.451	0.510	0.060	13.304

Table A2.1 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
170	c_{75}	0.523	0.584	0.061	11.663
150	c_{81}	0.522	0.545	0.023	4.406
150	c_{82}	0.496	0.489	-0.007	1.411
150	c_{83}	0.522	0.491	-0.031	5.939
150	c_{84}	0.543	0.545	0.002	0.368
150	c_{85}	0.781	0.619	-0.162	20.743
100	c_{91}	0.695	0.647	-0.048	6.906
100	c_{92}	0.682	0.618	-0.064	9.384
100	c_{93}	0.641	0.646	0.005	0.780
100	c_{94}	0.726	0.715	-0.011	1.515
100	c_{95}	0.758	0.821	0.063	8.311
100	c_{101}	0.634	0.642	0.008	1.262
100	c_{102}	0.662	0.618	-0.044	6.647
100	c_{103}	0.645	0.643	-0.002	0.310
100	c_{104}	0.717	0.715	-0.001	0.139
100	c_{105}	0.899	0.842	-0.057	6.340

Table A2.2
Constant Detection (d=2), $\bar{n}_j$ =500, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
500	c_{11}	0.341	0.354	0.014	4.106
500	c_{12}	0.330	0.321	-0.009	2.727
500	c_{13}	0.301	0.294	-0.007	2.326
500	c_{14}	0.286	0.284	-0.002	0.699
500	c_{15}	0.279	0.294	0.015	5.376
150	c_{21}	0.641	0.672	0.031	4.836
150	c_{22}	0.599	0.583	-0.016	2.671
150	c_{23}	0.548	0.527	-0.021	3.832
150	c_{24}	0.493	0.510	0.017	3.448
150	c_{25}	0.575	0.530	-0.045	7.826
500	c_{31}	0.358	0.342	-0.016	4.469
500	c_{32}	0.308	0.297	-0.011	3.571
500	c_{33}	0.296	0.267	-0.029	9.797
500	c_{34}	0.282	0.266	-0.016	5.674
500	c_{35}	0.319	0.298	-0.021	6.583
750	c_{41}	0.258	0.279	0.021	8.140
750	c_{42}	0.241	0.243	0.001	0.415
750	c_{43}	0.218	0.217	-0.001	0.459
750	c_{44}	0.197	0.217	0.020	10.152
750	c_{45}	0.238	0.243	0.004	1.681
250	c_{51}	0.496	0.455	-0.041	8.266
250	c_{52}	0.432	0.413	-0.019	4.398
250	c_{53}	0.399	0.398	-0.001	0.251
250	c_{54}	0.367	0.412	0.045	12.262
250	c_{55}	0.442	0.454	0.012	2.715
250	c_{61}	0.406	0.454	0.048	11.823
250	c_{62}	0.379	0.414	0.035	9.235
250	c_{63}	0.403	0.398	-0.006	1.489
250	c_{64}	0.384	0.413	0.029	7.552
250	c_{65}	0.449	0.454	0.005	1.114
850	c_{71}	0.226	0.228	0.002	0.885
850	c_{72}	0.208	0.204	-0.004	1.923
850	c_{73}	0.189	0.204	0.015	7.937
850	c_{74}	0.233	0.228	-0.005	2.146
850	c_{75}	0.267	0.262	-0.005	1.873
750	c_{81}	0.229	0.243	0.014	6.114
750	c_{82}	0.229	0.217	-0.012	5.240
750	c_{83}	0.198	0.217	0.019	9.596
750	c_{84}	0.241	0.243	0.002	0.830
750	c_{85}	0.307	0.279	-0.028	9.121

Table A2.2 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =500, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
500	c_{91}	0.273	0.293	0.020	7.326
500	c_{92}	0.285	0.284	-0.002	0.702
500	c_{93}	0.283	0.294	0.010	3.534
500	c_{94}	0.315	0.322	0.007	2.222
500	c_{95}	0.358	0.352	-0.006	1.676
500	c_{101}	0.290	0.293	0.003	1.034
500	c_{102}	0.299	0.283	-0.016	5.351
500	c_{103}	0.319	0.295	-0.025	7.837
500	c_{104}	0.337	0.321	-0.016	4.748
500	c_{105}	0.375	0.354	-0.021	5.600

Table A2.3
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
100	c_{11}	0.801	0.795	-0.006	0.749
100	c_{12}	0.787	0.736	-0.052	6.607
100	c_{13}	0.723	0.667	-0.056	7.746
100	c_{14}	0.652	0.619	-0.034	5.215
100	c_{15}	0.572	0.621	0.048	8.392
30	c_{21}	1.193	1.086	-0.108	9.053
30	c_{22}	1.147	1.111	-0.036	3.139
30	c_{23}	1.202	1.202	0.000	0.000
30	c_{24}	1.280	1.300	0.021	1.641
30	c_{25}	0.843	1.375	0.532	63.108
100	c_{31}	0.713	0.822	0.109	15.288
100	c_{32}	0.722	0.715	-0.007	0.970
100	c_{33}	0.662	0.647	-0.014	2.115
100	c_{34}	0.597	0.619	0.022	3.685
100	c_{35}	0.583	0.646	0.063	10.806
150	c_{41}	0.555	0.532	-0.023	4.144
150	c_{42}	0.551	0.509	-0.042	7.623
150	c_{43}	0.565	0.532	-0.033	5.841
150	c_{44}	0.685	0.583	-0.102	14.891
150	c_{45}	0.730	0.677	-0.053	7.260
50	c_{51}	0.798	1.126	0.327	40.977
50	c_{52}	1.147	0.976	-0.171	14.908
50	c_{53}	0.988	0.870	-0.119	12.045
50	c_{54}	0.883	0.854	-0.030	3.398
50	c_{55}	1.001	0.941	-0.061	6.094
50	c_{61}	0.981	0.937	-0.043	4.383
50	c_{62}	0.893	0.855	-0.038	4.255
50	c_{63}	0.815	0.868	0.054	6.626
50	c_{64}	1.055	0.977	-0.079	7.488
50	c_{65}	0.847	1.124	0.277	32.704
170	c_{71}	0.659	0.602	-0.057	8.649
170	c_{72}	0.498	0.514	0.017	3.414
170	c_{73}	0.426	0.459	0.032	7.512
170	c_{74}	0.426	0.459	0.032	7.512
170	c_{75}	0.514	0.514	0.000	0.000
150	c_{81}	0.509	0.547	0.038	7.466
150	c_{82}	0.476	0.489	0.013	2.731
150	c_{83}	0.492	0.489	-0.003	0.610
150	c_{84}	0.581	0.550	-0.031	5.336
150	c_{85}	0.681	0.636	-0.045	6.608

Table A2.3 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
100	c_{91}	0.772	0.770	-0.002	0.259
100	c_{92}	0.714	0.665	-0.049	6.863
100	c_{93}	0.618	0.601	-0.016	2.589
100	c_{94}	0.682	0.610	-0.072	10.557
100	c_{95}	0.701	0.682	-0.019	2.710
100	c_{101}	0.660	0.686	0.026	3.939
100	c_{102}	0.657	0.610	-0.047	7.154
100	c_{103}	0.668	0.604	-0.063	9.431
100	c_{104}	0.694	0.668	-0.026	3.746
100	c_{105}	0.750	0.776	0.026	3.467

Table A2.4
Varied Detection (d=1 to 5), $\bar{n}_j$ =500, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
500	c_{11}	0.311	0.344	0.033	10.611
500	c_{12}	0.338	0.340	0.003	0.888
500	c_{13}	0.308	0.297	-0.011	3.571
500	c_{14}	0.274	0.276	0.001	0.365
500	c_{15}	0.296	0.283	-0.013	4.392
150	c_{21}	0.542	0.509	-0.033	6.089
150	c_{22}	0.460	0.505	0.045	9.783
150	c_{23}	0.489	0.543	0.054	11.043
150	c_{24}	0.651	0.606	-0.045	6.912
150	c_{25}	0.724	0.656	-0.068	9.392
500	c_{31}	0.341	0.352	0.011	3.226
500	c_{32}	0.299	0.323	0.023	7.692
500	c_{33}	0.297	0.295	-0.002	0.673
500	c_{34}	0.274	0.284	0.010	3.650
500	c_{35}	0.277	0.293	0.016	5.776
750	c_{41}	0.233	0.239	0.006	2.575
750	c_{42}	0.233	0.232	-0.001	0.429
750	c_{43}	0.215	0.241	0.026	12.093
750	c_{44}	0.255	0.264	0.008	3.137
750	c_{45}	0.303	0.289	-0.014	4.620
250	c_{51}	0.538	0.497	-0.041	7.621
250	c_{52}	0.446	0.432	-0.014	3.139
250	c_{53}	0.398	0.387	-0.011	2.764
250	c_{54}	0.376	0.383	0.007	1.862
250	c_{55}	0.438	0.423	-0.015	3.425
250	c_{61}	0.355	0.422	0.067	18.873
250	c_{62}	0.306	0.383	0.077	25.163
250	c_{63}	0.372	0.387	0.014	3.763
250	c_{64}	0.404	0.433	0.029	7.178
250	c_{65}	0.533	0.498	-0.035	6.567
850	c_{71}	0.278	0.260	-0.018	6.475
850	c_{72}	0.215	0.228	0.013	6.047
850	c_{73}	0.203	0.204	0.001	0.493
850	c_{74}	0.185	0.204	0.019	10.270
850	c_{75}	0.212	0.228	0.015	7.075
750	c_{81}	0.238	0.243	0.005	2.101
750	c_{82}	0.236	0.217	-0.019	8.051
750	c_{83}	0.218	0.217	-0.001	0.459
750	c_{84}	0.219	0.242	0.024	10.959
750	c_{85}	0.234	0.276	0.042	17.949

Table A2.4 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =500, PME with Parameter Constraints

Sample size	Parameter	SD	MeanSE	Bias	%Bias
500	c_{91}	0.337	0.339	0.002	0.593
500	c_{92}	0.290	0.297	0.007	2.414
500	c_{93}	0.282	0.268	-0.014	4.965
500	c_{94}	0.263	0.270	0.007	2.662
500	c_{95}	0.308	0.304	-0.004	1.299
500	c_{101}	0.281	0.304	0.023	8.185
500	c_{102}	0.286	0.271	-0.015	5.245
500	c_{103}	0.255	0.269	0.014	5.490
500	c_{104}	0.254	0.297	0.043	16.929
500	c_{105}	0.322	0.339	0.016	4.969

Appendix B1

Parameter Estimates, Bias, Percent Bias, and MSE for Informative Priors (via MCMC) in Simulation 1 (Single Rater per Examinee)

Table B1.1
Constant Detection (d=2), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
100	d_1	2	4.270	2.270	113.500	5.197
30	d_2	2	4.260	2.260	113.000	5.232
100	d_3	2	4.145	2.145	107.250	4.653
150	d_4	2	4.132	2.132	106.600	4.583
50	d_5	2	3.911	1.911	95.550	3.722
50	d_6	2	3.850	1.850	92.500	3.501
170	d_7	2	3.694	1.694	84.700	2.901
150	d_8	2	3.667	1.667	83.350	2.834
100	d_9	2	3.260	1.260	63.000	1.646
100	d_{10}	2	3.242	1.242	62.100	1.601
100	c_{11}	-1	-0.813	0.187	18.700	0.575
100	c_{12}	1	1.993	0.993	99.300	1.600
100	c_{13}	3	5.419	2.419	80.630	6.616
100	c_{14}	5	9.209	4.209	84.180	18.434
100	c_{15}	7	13.040	6.040	86.290	37.076
30	c_{21}	-1	-0.502	0.498	49.800	0.870
30	c_{22}	1	2.262	1.262	126.200	2.625
30	c_{23}	3	5.593	2.593	86.430	8.092
30	c_{24}	5	9.115	4.115	82.300	18.105
30	c_{25}	7	12.993	5.993	85.610	36.717
100	c_{31}	0	0.115	0.115	.	0.528
100	c_{32}	2	3.285	1.285	64.250	2.279
100	c_{33}	4	6.708	2.708	67.700	7.949
100	c_{34}	6	10.602	4.602	76.700	21.863
100	c_{35}	8	14.640	6.640	83.000	44.600
150	c_{41}	0	0.008	0.008	.	0.270
150	c_{42}	2	3.119	1.119	55.950	1.589
150	c_{43}	4	6.874	2.874	71.850	8.606
150	c_{44}	6	10.772	4.772	79.530	23.198
150	c_{45}	8	14.694	6.694	83.680	45.121
50	c_{51}	1	1.338	0.338	33.800	0.706
50	c_{52}	3	4.665	1.665	55.500	3.546
50	c_{53}	5	8.211	3.211	64.220	11.036
50	c_{54}	7	11.912	4.912	70.170	24.779

Table B1.1 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
50	c_{55}	9	15.718	6.718	74.640	45.364
50	c_{61}	1	1.352	0.352	35.200	0.574
50	c_{62}	3	4.575	1.575	52.500	3.074
50	c_{63}	5	8.219	3.219	64.380	11.040
50	c_{64}	7	11.827	4.827	68.960	23.780
50	c_{65}	9	15.750	6.750	75.000	45.783
170	c_{71}	2	2.567	0.567	28.350	0.544
170	c_{72}	4	5.943	1.943	48.580	4.088
170	c_{73}	6	9.456	3.456	57.600	12.309
170	c_{74}	8	13.145	5.145	64.310	26.794
170	c_{75}	10	16.947	6.947	69.470	48.413
150	c_{81}	2	2.542	0.542	27.100	0.516
150	c_{82}	4	5.804	1.804	45.100	3.546
150	c_{83}	6	9.306	3.306	55.100	11.277
150	c_{84}	8	13.007	5.007	62.590	25.425
150	c_{85}	10	16.795	6.795	67.950	46.351
100	c_{91}	3	3.344	0.344	11.470	0.377
100	c_{92}	5	6.650	1.650	33.000	3.007
100	c_{93}	7	9.988	2.988	42.690	9.184
100	c_{94}	9	13.451	4.451	49.460	20.011
100	c_{95}	11	16.668	5.668	51.530	32.378
100	c_{101}	3	3.425	0.425	14.170	0.452
100	c_{102}	5	6.627	1.627	32.540	3.029
100	c_{103}	7	9.916	2.916	41.660	8.838
100	c_{104}	9	13.400	4.400	48.890	19.573
100	c_{105}	11	16.705	5.705	51.860	32.840
		Latent Class Sizes				
	Class 1	0.08	0.137	0.057	71.250	
	Class 2	0.17	0.220	0.050	29.410	
	Class 3	0.25	0.242	-0.008	3.200	
	Class 4	0.25	0.212	-0.038	15.200	
	Class 5	0.17	0.102	-0.068	40.000	
	Class 6	0.08	0.088	0.008	10.000	

Table B1.2
Constant Detection (d=2), $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
500	d_1	2	4.221	2.221	111.050	4.968
150	d_2	2	4.254	2.254	112.700	5.129
500	d_3	2	4.231	2.231	111.550	5.014
750	d_4	2	4.217	2.217	110.850	4.950
250	d_5	2	3.979	1.979	98.950	3.948
250	d_6	2	3.997	1.997	99.850	4.021
850	d_7	2	3.736	1.736	86.800	3.109
750	d_8	2	3.740	1.740	87.000	3.088
500	d_9	2	3.321	1.321	66.050	1.818
500	d_{10}	2	3.327	1.327	66.350	1.809
500	c_{11}	-1	-1.042	-0.042	4.200	0.177
500	c_{12}	1	1.752	0.752	75.200	0.935
500	c_{13}	3	5.135	2.135	71.170	4.950
500	c_{14}	5	9.075	4.075	81.500	17.094
500	c_{15}	7	13.079	6.079	86.840	37.491
150	c_{21}	-1	-0.918	0.082	8.200	0.410
150	c_{22}	1	1.921	0.921	92.100	1.353
150	c_{23}	3	5.374	2.374	79.130	6.300
150	c_{24}	5	9.261	4.261	85.220	18.723
150	c_{25}	7	13.161	6.161	88.010	38.496
500	c_{31}	0	0.208	0.208	.	0.291
500	c_{32}	2	3.384	1.384	69.200	2.260
500	c_{33}	4	7.101	3.101	77.530	10.018
500	c_{34}	6	11.048	5.048	84.130	25.906
500	c_{35}	8	15.140	7.140	89.250	51.442
750	c_{41}	0	0.123	0.123	.	0.186
750	c_{42}	2	3.344	1.344	67.200	2.027
750	c_{43}	4	6.975	2.975	74.380	9.098
750	c_{44}	6	10.954	4.954	82.570	24.886
750	c_{45}	8	15.029	7.029	87.860	49.819
250	c_{51}	1	1.499	0.499	49.900	0.544
250	c_{52}	3	4.787	1.787	59.570	3.568
250	c_{53}	5	8.484	3.484	69.680	12.551
250	c_{54}	7	12.242	5.242	74.890	27.871
250	c_{55}	9	16.326	7.326	81.400	53.948
250	c_{61}	1	1.525	0.525	52.500	0.531
250	c_{62}	3	4.818	1.818	60.600	3.612
250	c_{63}	5	8.537	3.537	70.740	12.914
250	c_{64}	7	12.356	5.356	76.510	29.044
250	c_{65}	9	16.394	7.394	82.160	54.979

Table B1.2 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
850	c_{71}	2	2.839	0.839	41.950	0.951
850	c_{72}	4	6.210	2.210	55.250	5.283
850	c_{73}	6	9.786	3.786	63.100	15.039
850	c_{74}	8	13.437	5.437	67.960	30.677
850	c_{75}	10	17.293	7.293	72.930	54.703
750	c_{81}	2	2.822	0.822	41.100	0.845
750	c_{82}	4	6.171	2.171	54.280	5.011
750	c_{83}	6	9.775	3.775	62.920	14.677
750	c_{84}	8	13.432	5.432	67.900	30.169
750	c_{85}	10	17.258	7.258	72.580	53.596
500	c_{91}	3	3.867	0.867	28.900	0.986
500	c_{92}	5	7.047	2.047	40.940	4.628
500	c_{93}	7	10.293	3.293	47.040	11.553
500	c_{94}	9	13.801	4.801	53.340	24.064
500	c_{95}	11	16.911	5.911	53.740	36.120
500	c_{101}	3	3.897	0.897	29.900	1.011
500	c_{102}	5	7.088	2.088	41.760	4.688
500	c_{103}	7	10.376	3.376	48.230	11.903
500	c_{104}	9	13.780	4.780	53.110	23.497
500	c_{105}	11	16.979	5.979	54.350	36.565
			Latent Class Sizes			
	Class 1	0.08	0.127	0.047	58.750	
	Class 2	0.17	0.221	0.051	30.000	
	Class 3	0.25	0.240	-0.010	4.000	
	Class 4	0.25	0.219	-0.031	12.400	
	Class 5	0.17	0.110	-0.060	35.290	
	Class 6	0.08	0.083	0.003	3.750	

Table B1.3
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
100	d_1	1.0	5.177	4.177	417.700	17.505
30	d_2	1.0	2.664	1.664	166.400	2.967
100	d_3	2.0	4.314	2.314	115.700	5.419
150	d_4	2.0	3.182	1.182	59.100	1.437
50	d_5	3.0	4.024	1.024	34.130	1.131
50	d_6	3.0	3.259	0.259	8.630	0.148
170	d_7	4.0	3.931	-0.069	1.730	0.054
150	d_8	4.0	3.440	-0.560	14.000	0.354
100	d_9	5.0	3.919	-1.081	21.620	1.228
100	d_{10}	5.0	3.486	-1.514	30.280	2.334
100	c_{11}	-1.5	-0.123	1.377	91.800	2.503
100	c_{12}	-0.5	2.238	2.738	547.600	8.274
100	c_{13}	0.5	4.842	4.342	868.400	19.737
100	c_{14}	1.5	7.931	6.431	428.730	42.398
100	c_{15}	2.5	11.518	9.018	360.720	82.185
30	c_{21}	2.5	4.417	1.917	76.680	3.886
30	c_{22}	3.5	6.971	3.471	99.170	12.483
30	c_{23}	4.5	9.445	4.945	109.890	25.145
30	c_{24}	5.5	11.877	6.377	115.950	41.653
30	c_{25}	6.5	14.660	8.160	125.540	68.272
100	c_{31}	-1.0	-0.506	0.494	49.400	0.686
100	c_{32}	1.0	2.359	1.359	135.900	2.389
100	c_{33}	3.0	5.581	2.581	86.030	7.398
100	c_{34}	5.0	9.390	4.390	87.800	19.954
100	c_{35}	7.0	13.365	6.365	90.930	40.995
150	c_{41}	3.0	3.660	0.660	22.000	0.661
150	c_{42}	5.0	6.729	1.729	34.580	3.253
150	c_{43}	7.0	10.011	3.011	43.010	9.371
150	c_{44}	9.0	13.301	4.301	47.790	18.799
150	c_{45}	11.0	16.410	5.410	49.180	29.558
50	c_{51}	-0.5	-0.284	0.216	43.200	0.589
50	c_{52}	2.5	2.822	0.322	12.880	0.946
50	c_{53}	5.5	6.315	0.815	14.820	1.632
50	c_{54}	8.5	10.082	1.582	18.610	3.350
50	c_{55}	11.5	14.082	2.582	22.450	7.184
50	c_{61}	3.5	2.634	-0.866	24.740	1.117
50	c_{62}	6.5	5.924	-0.576	8.860	0.786
50	c_{63}	9.5	9.337	-0.163	1.720	0.513
50	c_{64}	12.5	12.934	0.434	3.470	0.616
50	c_{65}	15.5	16.576	1.076	6.940	1.403

Table B1.3 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
170	c_{71}	0.0	-0.250	-0.250	.	0.405
170	c_{72}	4.0	3.008	-0.992	24.800	1.367
170	c_{73}	8.0	6.570	-1.430	17.880	2.542
170	c_{74}	12.0	10.579	-1.421	11.840	2.476
170	c_{75}	16.0	14.728	-1.272	7.950	1.900
150	c_{81}	4.0	2.384	-1.616	40.400	2.810
150	c_{82}	8.0	5.655	-2.345	29.310	5.794
150	c_{83}	12.0	9.261	-2.739	22.830	7.971
150	c_{84}	16.0	13.113	-2.887	18.040	8.726
150	c_{85}	20.0	16.840	-3.160	15.800	10.216
100	c_{91}	0.5	0.213	-0.287	57.400	0.501
100	c_{92}	5.5	3.506	-1.994	36.250	4.518
100	c_{93}	10.5	6.987	-3.513	33.460	12.874
100	c_{94}	15.5	10.938	-4.562	29.430	21.369
100	c_{95}	20.5	14.998	-5.502	26.840	30.647
100	c_{101}	4.5	2.033	-2.467	54.820	6.404
100	c_{102}	9.5	5.317	-4.183	44.030	17.929
100	c_{103}	14.5	8.894	-5.606	38.660	31.909
100	c_{104}	19.5	12.729	-6.771	34.720	46.276
100	c_{105}	24.5	16.670	-7.830	31.960	61.427
			Latent Class Sizes			
	Class 1	0.08	0.108	0.028	35.000	
	Class 2	0.17	0.230	0.060	35.290	
	Class 3	0.25	0.246	-0.004	1.600	
	Class 4	0.25	0.218	-0.032	12.800	
	Class 5	0.17	0.105	-0.065	38.240	
	Class 6	0.08	0.092	0.012	15.000	

Table B1.4
Varied Detection (d=1 to 5), $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
500	d_1	1.0	5.270	4.270	427.000	18.274
150	d_2	1.0	2.628	1.628	162.800	2.727
500	d_3	2.0	4.356	2.356	117.800	5.597
750	d_4	2.0	3.261	1.261	63.050	1.680
250	d_5	3.0	4.198	1.198	39.930	1.479
250	d_6	3.0	3.408	0.408	13.600	0.212
850	d_7	4.0	4.129	0.129	3.230	0.075
750	d_8	4.0	3.530	-0.470	11.750	0.275
500	d_9	5.0	3.986	-1.014	20.280	1.081
500	d_{10}	5.0	3.563	-1.437	28.740	2.113
500	c_{11}	-1.5	-0.198	1.302	86.800	2.060
500	c_{12}	-0.5	2.030	2.530	506.000	7.080
500	c_{13}	0.5	4.637	4.137	827.400	17.615
500	c_{14}	1.5	7.730	6.230	415.330	39.481
500	c_{15}	2.5	11.308	8.808	352.320	78.201
150	c_{21}	2.5	5.237	2.737	109.480	7.625
150	c_{22}	3.5	7.398	3.898	111.370	15.489
150	c_{23}	4.5	9.478	4.978	110.620	25.267
150	c_{24}	5.5	11.601	6.101	110.930	37.937
150	c_{25}	6.5	13.634	7.134	109.750	51.944
500	c_{31}	-1.0	-0.568	0.432	43.200	0.470
500	c_{32}	1.0	2.329	1.329	132.900	2.220
500	c_{33}	3.0	5.447	2.447	81.570	6.516
500	c_{34}	5.0	9.321	4.321	86.420	19.262
500	c_{35}	7.0	13.324	6.324	90.340	40.606
750	c_{41}	3.0	4.008	1.008	33.600	1.305
750	c_{42}	5.0	6.969	1.969	39.380	4.384
750	c_{43}	7.0	10.113	3.113	44.470	10.530
750	c_{44}	9.0	13.594	4.594	51.040	22.421
750	c_{45}	11.0	16.670	5.670	51.550	33.763
250	c_{51}	-0.5	-0.303	0.197	39.400	0.512
250	c_{52}	2.5	2.899	0.399	15.960	0.590
250	c_{53}	5.5	6.339	0.839	15.250	1.182
250	c_{54}	8.5	10.347	1.847	21.730	3.901
250	c_{55}	11.5	14.513	3.013	26.200	9.509
250	c_{61}	3.5	2.978	-0.522	14.910	0.446
250	c_{62}	6.5	6.193	-0.307	4.720	0.308
250	c_{63}	9.5	9.634	0.134	1.410	0.360
250	c_{64}	12.5	13.382	0.882	7.060	1.209
250	c_{65}	15.5	17.001	1.501	9.680	2.663

Table B1.4 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
850	c_{71}	0.0	-0.111	-0.111	.	0.361
850	c_{72}	4.0	3.235	-0.765	19.130	0.962
850	c_{73}	8.0	6.828	-1.172	14.650	1.880
850	c_{74}	12.0	10.932	-1.068	8.900	1.708
850	c_{75}	16.0	15.220	-0.780	4.880	1.316
750	c_{81}	4.0	2.586	-1.414	35.350	2.238
750	c_{82}	8.0	5.770	-2.230	27.880	5.313
750	c_{83}	12.0	9.352	-2.648	22.070	7.515
750	c_{84}	16.0	13.168	-2.832	17.700	8.704
750	c_{85}	20.0	17.227	-2.773	13.870	8.560
500	c_{91}	0.5	0.183	-0.317	63.400	0.396
500	c_{92}	5.5	3.461	-2.039	37.070	4.459
500	c_{93}	10.5	7.080	-3.420	32.570	12.089
500	c_{94}	15.5	11.103	-4.397	28.370	19.744
500	c_{95}	20.5	15.412	-5.088	24.820	26.366
500	c_{101}	4.5	2.210	-2.290	50.890	5.465
500	c_{102}	9.5	5.366	-4.134	43.520	17.353
500	c_{103}	14.5	8.934	-5.566	38.390	31.363
500	c_{104}	19.5	12.845	-6.655	34.130	44.847
500	c_{105}	24.5	17.033	-7.467	30.480	56.440
			Latent Class Sizes			
	Class 1	0.08	0.099	0.019	23.750	
	Class 2	0.17	0.242	0.072	42.350	
	Class 3	0.25	0.248	-0.002	0.800	
	Class 4	0.25	0.224	-0.026	10.400	
	Class 5	0.17	0.101	-0.069	40.590	
	Class 6	0.08	0.084	0.004	5.000	

Appendix B2

Evaluation of the Estimated Posterior Standard Deviations for Informative Priors (via MCMC) in Simulation 1 (Single Rater per Examinee)

Table B2.1
Constant Detection (d=2), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
100	d_1	0.210	1.036	0.826	393.333
30	d_2	0.355	1.114	0.759	213.803
100	d_3	0.231	0.984	0.753	325.974
150	d_4	0.193	0.966	0.773	400.518
50	d_5	0.267	0.960	0.693	259.551
50	d_6	0.283	0.943	0.661	233.569
170	d_7	0.179	0.858	0.679	379.330
150	d_8	0.237	0.865	0.627	264.557
100	d_9	0.242	0.816	0.574	237.190
100	d_{10}	0.243	0.814	0.571	234.979
100	c_{11}	0.738	0.950	0.211	28.591
100	c_{12}	0.787	1.337	0.550	69.886
100	c_{13}	0.879	1.783	0.904	102.844
100	c_{14}	0.850	2.408	1.558	183.294
100	c_{15}	0.775	3.044	2.269	292.774
30	c_{21}	0.792	1.395	0.602	76.010
30	c_{22}	1.021	1.703	0.682	66.797
30	c_{23}	1.176	2.121	0.945	80.357
30	c_{24}	1.086	2.648	1.562	143.831
30	c_{25}	0.902	3.209	2.306	255.654
100	c_{31}	0.721	1.015	0.294	40.777
100	c_{32}	0.796	1.397	0.601	75.503
100	c_{33}	0.789	1.944	1.155	146.388
100	c_{34}	0.834	2.573	1.739	208.513
100	c_{35}	0.718	3.216	2.498	347.911
150	c_{41}	0.522	0.898	0.376	72.031
150	c_{42}	0.582	1.307	0.725	124.570
150	c_{43}	0.593	1.880	1.287	217.032
150	c_{44}	0.656	2.528	1.872	285.366
150	c_{45}	0.556	3.178	2.622	471.583
50	c_{51}	0.773	1.198	0.425	54.981
50	c_{52}	0.883	1.634	0.751	85.051
50	c_{53}	0.856	2.216	1.360	158.879
50	c_{54}	0.811	2.823	2.013	248.212
50	c_{55}	0.489	3.419	2.930	599.182

Table B2.1 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
50	c_{61}	0.674	1.177	0.503	74.629
50	c_{62}	0.774	1.623	0.849	109.690
50	c_{63}	0.828	2.209	1.381	166.787
50	c_{64}	0.697	2.813	2.116	303.587
50	c_{65}	0.475	3.405	2.929	616.632
170	c_{71}	0.474	1.027	0.552	116.456
170	c_{72}	0.564	1.565	1.001	177.482
170	c_{73}	0.607	2.196	1.589	261.779
170	c_{74}	0.576	2.866	2.290	397.569
170	c_{75}	0.402	3.484	3.083	766.915
150	c_{81}	0.474	1.040	0.566	119.409
150	c_{82}	0.541	1.578	1.037	191.682
150	c_{83}	0.591	2.214	1.624	274.788
150	c_{84}	0.600	2.890	2.290	381.667
150	c_{85}	0.429	3.505	3.077	717.249
100	c_{91}	0.511	1.127	0.615	120.352
100	c_{92}	0.535	1.751	1.216	227.290
100	c_{93}	0.506	2.403	1.898	375.099
100	c_{94}	0.452	3.067	2.615	578.540
100	c_{95}	0.506	3.534	3.028	598.419
100	c_{101}	0.524	1.143	0.619	118.130
100	c_{102}	0.621	1.758	1.137	183.092
100	c_{103}	0.581	2.407	1.826	314.286
100	c_{104}	0.463	3.087	2.623	566.523
100	c_{105}	0.544	3.553	3.009	553.125
	Latent Class Sizes				
	Class 1	0.018	0.043	0.026	144.444
	Class 2	0.021	0.059	0.037	176.190
	Class 3	0.018	0.063	0.045	250.000
	Class 4	0.017	0.060	0.043	252.941
	Class 5	0.016	0.053	0.037	231.250
	Class 6	0.011	0.047	0.035	318.182

Table B2.2
Constant Detection (d=2), $\bar{n}_j$ =500, MCM with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
500	d_1	0.188	0.969	0.782	415.957
150	d_2	0.224	1.016	0.791	353.125
500	d_3	0.196	0.942	0.746	380.612
750	d_4	0.192	0.937	0.745	388.021
250	d_5	0.173	0.895	0.722	417.341
250	d_6	0.187	0.888	0.701	374.866
850	d_7	0.313	0.830	0.517	165.176
750	d_8	0.245	0.826	0.581	237.143
500	d_9	0.272	0.761	0.489	179.779
500	d_{10}	0.218	0.775	0.556	255.046
500	c_{11}	0.421	0.602	0.182	43.230
500	c_{12}	0.610	1.101	0.491	80.492
500	c_{13}	0.630	1.528	0.898	142.540
500	c_{14}	0.705	2.254	1.549	219.716
500	c_{15}	0.733	2.958	2.225	303.547
150	c_{21}	0.638	0.843	0.205	32.132
150	c_{22}	0.714	1.244	0.530	74.230
150	c_{23}	0.818	1.689	0.871	106.479
150	c_{24}	0.759	2.376	1.618	213.175
150	c_{25}	0.741	3.046	2.305	311.066
500	c_{31}	0.500	0.771	0.271	54.200
500	c_{32}	0.589	1.178	0.589	100.000
500	c_{33}	0.635	1.795	1.159	182.520
500	c_{34}	0.651	2.477	1.825	280.338
500	c_{35}	0.679	3.152	2.472	364.065
750	c_{41}	0.416	0.726	0.310	74.519
750	c_{42}	0.474	1.149	0.675	142.405
750	c_{43}	0.498	1.763	1.264	253.815
750	c_{44}	0.585	2.457	1.872	320.000
750	c_{45}	0.651	3.144	2.494	383.103
250	c_{51}	0.546	0.964	0.418	76.557
250	c_{52}	0.614	1.381	0.766	124.756
250	c_{53}	0.644	2.040	1.396	216.770
250	c_{54}	0.630	2.695	2.065	327.778
250	c_{55}	0.529	3.394	2.865	541.588
250	c_{61}	0.508	0.972	0.463	91.142
250	c_{62}	0.557	1.380	0.823	147.756
250	c_{63}	0.637	2.036	1.398	219.466
250	c_{64}	0.605	2.683	2.078	343.471
250	c_{65}	0.555	3.369	2.815	507.207

Table B2.2 (Continued)
Constant Detection (d=2), $\bar{n}_j$ =500, MCM with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
850	c_{71}	0.499	0.932	0.433	86.774
850	c_{72}	0.636	1.509	0.873	137.264
850	c_{73}	0.843	2.169	1.326	157.295
850	c_{74}	1.063	2.840	1.776	167.074
850	c_{75}	1.238	3.510	2.272	183.522
750	c_{81}	0.413	0.918	0.505	122.276
750	c_{82}	0.550	1.483	0.933	169.636
750	c_{83}	0.658	2.147	1.488	226.140
750	c_{84}	0.818	2.811	1.993	243.643
750	c_{85}	0.964	3.465	2.501	259.440
500	c_{91}	0.487	1.044	0.556	114.168
500	c_{92}	0.666	1.676	1.010	151.652
500	c_{93}	0.846	2.294	1.448	171.158
500	c_{94}	1.013	2.973	1.960	193.485
500	c_{95}	1.092	3.448	2.356	215.751
500	c_{101}	0.457	1.031	0.574	125.602
500	c_{102}	0.578	1.662	1.084	187.543
500	c_{103}	0.716	2.297	1.580	220.670
500	c_{104}	0.813	2.986	2.174	267.405
500	c_{105}	0.910	3.486	2.576	283.077
	Latent Class Sizes				
	Class 1	0.019	0.039	0.020	105.263
	Class 2	0.014	0.050	0.036	257.143
	Class 3	0.013	0.054	0.041	315.385
	Class 4	0.013	0.050	0.038	292.308
	Class 5	0.011	0.047	0.036	327.273
	Class 6	0.011	0.040	0.030	272.727

Table B2.3
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
100	d_1	0.235	1.199	0.964	410.213
30	d_2	0.448	0.870	0.422	94.196
100	d_3	0.260	1.052	0.791	304.231
150	d_4	0.199	0.806	0.607	305.025
50	d_5	0.287	1.017	0.730	254.355
50	d_6	0.285	0.850	0.565	198.246
170	d_7	0.224	0.941	0.717	320.089
150	d_8	0.202	0.832	0.629	311.386
100	d_9	0.244	0.953	0.709	290.574
100	d_{10}	0.203	0.850	0.646	318.227
100	c_{11}	0.783	1.147	0.365	46.616
100	c_{12}	0.886	1.556	0.669	75.508
100	c_{13}	0.945	1.761	0.816	86.349
100	c_{14}	1.022	2.326	1.303	127.495
100	c_{15}	0.928	2.852	1.923	207.220
30	c_{21}	0.462	1.431	0.969	209.740
30	c_{22}	0.662	2.066	1.403	211.934
30	c_{23}	0.836	2.670	1.834	219.378
30	c_{24}	0.999	3.190	2.191	219.319
30	c_{25}	1.308	3.607	2.298	175.688
100	c_{31}	0.668	1.070	0.402	60.180
100	c_{32}	0.739	1.414	0.675	91.340
100	c_{33}	0.862	1.829	0.967	112.181
100	c_{34}	0.827	2.466	1.639	198.186
100	c_{35}	0.694	3.110	2.416	348.127
150	c_{41}	0.478	1.144	0.666	139.331
150	c_{42}	0.515	1.764	1.249	242.524
150	c_{43}	0.554	2.425	1.871	337.726
150	c_{44}	0.547	3.091	2.543	464.899
150	c_{45}	0.538	3.547	3.009	559.294
50	c_{51}	0.740	1.254	0.514	69.459
50	c_{52}	0.922	1.531	0.609	66.052
50	c_{53}	0.989	2.030	1.041	105.258
50	c_{54}	0.926	2.643	1.717	185.421
50	c_{55}	0.722	3.269	2.546	352.632
50	c_{61}	0.609	1.185	0.575	94.417
50	c_{62}	0.677	1.756	1.079	159.380
50	c_{63}	0.701	2.404	1.703	242.939
50	c_{64}	0.657	3.073	2.416	367.732

Table B2.3 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
50	c_{65}	0.497	3.559	3.062	616.097
170	c_{71}	0.588	0.928	0.340	57.823
170	c_{72}	0.622	1.269	0.648	104.180
170	c_{73}	0.709	1.841	1.132	159.661
170	c_{74}	0.680	2.529	1.849	271.912
170	c_{75}	0.534	3.216	2.682	502.247
150	c_{81}	0.448	1.021	0.573	127.902
150	c_{82}	0.547	1.566	1.019	186.289
150	c_{83}	0.688	2.243	1.554	225.872
150	c_{84}	0.629	2.965	2.336	371.383
150	c_{85}	0.483	3.542	3.058	633.126
100	c_{91}	0.650	1.053	0.403	62.000
100	c_{92}	0.740	1.394	0.655	88.514
100	c_{93}	0.735	1.980	1.245	169.388
100	c_{94}	0.751	2.647	1.896	252.463
100	c_{95}	0.613	3.307	2.694	439.478
100	c_{101}	0.566	1.065	0.499	88.163
100	c_{102}	0.663	1.572	0.909	137.104
100	c_{103}	0.695	2.236	1.541	221.727
100	c_{104}	0.663	2.941	2.278	343.590
100	c_{105}	0.350	3.530	3.180	908.571
	Latent Class Sizes				
	Class 1	0.018	0.041	0.023	127.778
	Class 2	0.022	0.060	0.038	172.727
	Class 3	0.023	0.065	0.042	182.609
	Class 4	0.018	0.065	0.047	261.111
	Class 5	0.016	0.059	0.042	262.500
	Class 6	0.011	0.050	0.038	345.455

Table B2.4
Varied Detection (d=1 to 5), $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
500	d_1	0.201	1.165	0.963	479.104
150	d_2	0.277	0.774	0.498	179.783
500	d_3	0.217	0.995	0.778	358.525
750	d_4	0.302	0.766	0.464	153.642
250	d_5	0.208	0.973	0.765	367.788
250	d_6	0.214	0.816	0.602	281.308
850	d_7	0.244	0.926	0.682	279.508
750	d_8	0.233	0.808	0.575	246.781
500	d_9	0.232	0.922	0.690	297.414
500	d_{10}	0.219	0.812	0.593	270.776
500	c_{11}	0.607	0.935	0.327	53.871
500	c_{12}	0.827	1.467	0.640	77.388
500	c_{13}	0.712	1.537	0.825	115.871
500	c_{14}	0.819	2.156	1.337	163.248
500	c_{15}	0.793	2.696	1.903	239.975
150	c_{21}	0.364	1.424	1.060	291.209
150	c_{22}	0.543	1.948	1.405	258.748
150	c_{23}	0.698	2.456	1.758	251.862
150	c_{24}	0.847	2.978	2.131	251.594
150	c_{25}	1.029	3.341	2.312	224.684
500	c_{31}	0.535	0.793	0.258	48.224
500	c_{32}	0.676	1.241	0.564	83.432
500	c_{33}	0.731	1.604	0.873	119.425
500	c_{34}	0.774	2.314	1.540	198.966
500	c_{35}	0.783	3.021	2.238	285.824
750	c_{41}	0.540	1.029	0.489	90.556
750	c_{42}	0.716	1.627	0.912	127.374
750	c_{43}	0.921	2.253	1.332	144.625
750	c_{44}	1.153	2.947	1.794	155.594
750	c_{45}	1.277	3.427	2.150	168.363
250	c_{51}	0.691	0.909	0.218	31.548
250	c_{52}	0.660	1.248	0.589	89.242
250	c_{53}	0.695	1.783	1.087	156.403
250	c_{54}	0.703	2.487	1.784	253.770
250	c_{55}	0.662	3.168	2.506	378.550
250	c_{61}	0.419	0.980	0.561	133.890
250	c_{62}	0.465	1.557	1.093	235.054
250	c_{63}	0.588	2.218	1.631	277.381
250	c_{64}	0.659	2.935	2.276	345.372
250	c_{54}	0.703	2.487	1.784	253.770

Table B2.4 (Continued)
Varied Detection (d=1 to 5), $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
250	c_{55}	0.662	3.168	2.506	378.550
250	c_{61}	0.419	0.980	0.561	133.890
250	c_{62}	0.465	1.557	1.093	235.054
250	c_{63}	0.588	2.218	1.631	277.381
250	c_{64}	0.659	2.935	2.276	345.372
250	c_{65}	0.643	3.514	2.871	446.501
850	c_{71}	0.593	0.800	0.207	34.907
850	c_{72}	0.616	1.153	0.536	87.013
850	c_{73}	0.715	1.746	1.031	144.196
850	c_{74}	0.757	2.459	1.701	224.703
850	c_{75}	0.845	3.153	2.308	273.136
750	c_{81}	0.491	0.927	0.436	88.798
750	c_{82}	0.587	1.453	0.865	147.359
750	c_{83}	0.713	2.125	1.412	198.036
750	c_{84}	0.830	2.822	1.992	240.000
750	c_{85}	0.938	3.487	2.548	271.642
500	c_{91}	0.547	0.883	0.337	61.609
500	c_{92}	0.552	1.193	0.641	116.123
500	c_{93}	0.628	1.828	1.200	191.083
500	c_{94}	0.644	2.549	1.904	295.652
500	c_{95}	0.695	3.279	2.584	371.799
500	c_{101}	0.475	0.918	0.443	93.263
500	c_{102}	0.519	1.393	0.875	168.593
500	c_{103}	0.623	2.065	1.443	231.621
500	c_{104}	0.750	2.777	2.027	270.267
500	c_{105}	0.829	3.452	2.622	316.285
	Latent Class Sizes				
	Class 1	0.022	0.039	0.016	72.727
	Class 2	0.016	0.053	0.037	231.250
	Class 3	0.014	0.055	0.041	292.857
	Class 4	0.014	0.054	0.040	285.714
	Class 5	0.013	0.051	0.038	292.308
	Class 6	0.013	0.045	0.032	246.154

Appendix C1

Parameter Estimates, Bias, Percent Bias, and MSE for Bayes' Constants (via PME) in Simulation 2 (Partial Second Rater per Examinee)

Table C1.1
10% 2nd rater, $\bar{n}_j$ =100, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
110	d_1	1.0	0.820	-0.180	18.000	0.250
47	d_2	1.0	0.966	-0.034	3.400	0.219
110	d_3	2.0	1.310	-0.690	34.500	0.675
160	d_4	2.0	1.643	-0.357	17.850	0.305
53	d_5	3.0	0.748	-2.252	75.070	5.618
55	d_6	3.0	0.791	-2.209	73.630	5.433
185	d_7	4.0	1.806	-2.194	54.850	4.908
165	d_8	4.0	1.920	-2.080	52.000	4.445
110	d_9	5.0	1.813	-3.187	63.740	10.354
110	d_{10}	5.0	1.793	-3.207	64.140	10.517
110	c_{11}	-1.5	-2.051	-0.551	36.730	1.614
110	c_{12}	-0.5	-1.110	-0.610	122.000	1.080
110	c_{13}	0.5	-0.196	-0.696	139.200	1.175
110	c_{14}	1.5	0.772	-0.728	48.530	1.727
110	c_{15}	2.5	1.873	-0.627	25.080	2.471
47	c_{21}	2.5	2.174	-0.326	13.040	2.600
47	c_{22}	3.5	3.303	-0.197	5.630	4.312
47	c_{23}	4.5	4.409	-0.091	2.020	6.929
47	c_{24}	5.5	5.321	-0.179	3.250	9.176
47	c_{25}	6.5	6.189	-0.311	4.780	13.059
110	c_{31}	-1.0	-2.050	-1.050	105.000	2.351
110	c_{32}	1.0	-0.488	-1.488	148.800	2.923
110	c_{33}	3.0	1.092	-1.908	63.600	5.011
110	c_{34}	5.0	2.820	-2.180	43.600	8.433
110	c_{35}	7.0	4.610	-2.390	34.140	13.465
160	c_{41}	3.0	1.486	-1.514	50.470	3.553
160	c_{42}	5.0	3.637	-1.363	27.260	5.876
160	c_{43}	7.0	5.803	-1.197	17.100	10.339
160	c_{44}	9.0	7.671	-1.329	14.770	16.054
160	c_{45}	11.0	9.400	-1.600	14.550	22.540
53	c_{51}	-0.5	-2.517	-2.017	403.400	8.215
53	c_{52}	2.5	-1.009	-3.509	140.360	16.076
53	c_{53}	5.5	0.569	-4.931	89.650	27.802
53	c_{54}	8.5	2.093	-6.407	75.380	45.916
53	c_{55}	11.5	3.675	-7.825	68.040	68.305

Table C1.1 (Continued)
10% 2nd rater, $\bar{n}_j$ =100, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
55	c_{61}	3.5	-0.399	-3.899	111.400	18.464
55	c_{62}	6.5	1.211	-5.289	81.370	32.033
55	c_{63}	9.5	2.819	-6.681	70.330	50.632
55	c_{64}	12.5	4.364	-8.136	65.090	75.445
55	c_{65}	15.5	6.011	-9.489	61.220	102.310
185	c_{71}	0.0	-1.571	-1.571	.	3.007
185	c_{72}	4.0	0.403	-3.597	89.930	13.312
185	c_{73}	8.0	2.811	-5.189	64.860	29.181
185	c_{74}	12.0	5.336	-6.664	55.530	50.304
185	c_{75}	16.0	7.758	-8.242	51.510	80.115
165	c_{81}	4.0	0.440	-3.560	89.000	12.950
165	c_{82}	8.0	2.952	-5.048	63.100	28.101
165	c_{83}	12.0	5.531	-6.469	53.910	49.643
165	c_{84}	16.0	8.071	-7.929	49.560	78.887
165	c_{85}	20.0	10.083	-9.917	49.590	121.906
110	c_{91}	0.5	-1.368	-1.868	373.600	4.702
110	c_{92}	5.5	0.637	-4.863	88.420	24.957
110	c_{93}	10.5	3.025	-7.475	71.190	59.706
110	c_{94}	15.5	5.569	-9.931	64.070	106.996
110	c_{95}	20.5	7.906	-12.594	61.430	174.356
110	c_{101}	4.5	0.045	-4.455	99.000	20.784
110	c_{102}	9.5	2.382	-7.118	74.930	52.961
110	c_{103}	14.5	4.894	-9.606	66.250	98.624
110	c_{104}	19.5	7.364	-12.136	62.240	159.775
110	c_{105}	24.5	9.479	-15.021	61.310	246.122
			Latent Class Sizes			
	Class 1	0.08	0.241	0.161	201.250	
	Class 2	0.17	0.072	-0.098	57.650	
	Class 3	0.25	0.200	-0.050	20.000	
	Class 4	0.25	0.188	-0.062	24.800	
	Class 5	0.17	0.079	-0.091	53.530	
	Class 6	0.08	0.220	0.140	175.000	

Table C1.2
10% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
550	d_1	1.0	0.927	-0.073	7.300	0.099
235	d_2	1.0	0.884	-0.116	11.600	0.074
550	d_3	2.0	1.684	-0.316	15.800	0.262
800	d_4	2.0	1.759	-0.241	12.050	0.157
265	d_5	3.0	1.879	-1.121	37.367	1.399
275	d_6	3.0	1.936	-1.064	35.467	1.241
925	d_7	4.0	2.777	-1.223	30.575	1.666
825	d_8	4.0	2.800	-1.200	30.000	1.619
550	d_9	5.0	3.030	-1.970	39.400	4.039
550	d_{10}	5.0	3.045	-1.955	39.100	3.992
550	c_{11}	-1.5	-1.833	-0.333	22.200	0.567
550	c_{12}	-0.5	-0.869	-0.369	73.800	0.314
550	c_{13}	0.5	0.105	-0.395	79.000	0.322
550	c_{14}	1.5	1.125	-0.375	25.000	0.663
550	c_{15}	2.5	2.122	-0.378	15.120	1.36
235	c_{21}	2.5	2.043	-0.457	18.280	1.336
235	c_{22}	3.5	3.017	-0.483	13.800	2.297
235	c_{23}	4.5	3.953	-0.547	12.156	3.481
235	c_{24}	5.5	4.895	-0.605	11.000	5.036
235	c_{25}	6.5	5.883	-0.617	9.492	7.016
550	c_{31}	-1.0	-1.590	-0.590	59.000	0.843
550	c_{32}	1.0	0.106	-0.894	89.400	0.98
550	c_{33}	3.0	1.954	-1.046	34.867	2.113
550	c_{34}	5.0	3.874	-1.126	22.520	4.407
550	c_{35}	7.0	5.774	-1.226	17.514	8.105
800	c_{41}	3.0	2.126	-0.874	29.133	1.55
800	c_{42}	5.0	4.089	-0.911	18.220	3.372
800	c_{43}	7.0	6.083	-0.917	13.100	6.445
800	c_{44}	9.0	8.055	-0.945	10.500	10.418
800	c_{45}	11.0	9.851	-1.149	10.445	14.863
265	c_{51}	-0.5	-1.384	-0.884	176.800	1.296
265	c_{52}	2.5	0.619	-1.881	75.240	3.896
265	c_{53}	5.5	2.869	-2.631	47.836	8.68
265	c_{54}	8.5	5.078	-3.422	40.259	16.189
265	c_{55}	11.5	7.368	-4.132	35.930	26.16
275	c_{61}	3.5	1.418	-2.082	59.486	4.962
275	c_{62}	6.5	3.805	-2.695	41.462	9.79
275	c_{63}	9.5	6.086	-3.414	35.937	17.589
275	c_{64}	12.5	8.423	-4.077	32.616	27.776
275	c_{65}	15.5	10.494	-5.006	32.297	41.455

Table C1.2 (Continued)
10% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
925	c_{71}	0.0	-1.039	-1.039	.	1.377
925	c_{72}	4.0	1.586	-2.414	60.350	6.704
925	c_{73}	8.0	4.924	-3.076	38.450	13.431
925	c_{74}	12.0	7.990	-4.010	33.417	25.736
925	c_{75}	16.0	11.319	-4.681	29.256	41.423
825	c_{81}	4.0	1.583	-2.417	60.425	6.616
825	c_{82}	8.0	4.967	-3.033	37.913	13.232
825	c_{83}	12.0	8.059	-3.941	32.842	25.518
825	c_{84}	16.0	11.407	-4.593	28.706	41.066
825	c_{85}	20.0	14.093	-5.907	29.535	63.958
550	c_{91}	0.5	-0.736	-1.236	247.200	1.868
550	c_{92}	5.5	2.142	-3.358	61.055	12.666
550	c_{93}	10.5	5.792	-4.708	44.838	27.527
550	c_{94}	15.5	9.164	-6.336	40.877	52.322
550	c_{95}	20.5	12.790	-7.710	37.610	84.219
550	c_{101}	4.5	1.274	-3.226	71.689	11.133
550	c_{102}	9.5	4.947	-4.553	47.926	24.616
550	c_{103}	14.5	8.300	-6.200	42.759	48.263
550	c_{104}	19.5	12.004	-7.496	38.441	77.306
550	c_{105}	24.5	15.032	-9.468	38.645	121.213
			Latent Class Sizes			
	Class 1	0.08	0.163	0.083	103.750	
	Class 2	0.17	0.120	-0.050	29.412	
	Class 3	0.25	0.224	-0.026	10.400	
	Class 4	0.25	0.223	-0.027	10.800	
	Class 5	0.17	0.116	-0.054	31.765	
	Class 6	0.08	0.153	0.073	91.250	

Table C1.3
30% 2nd rater, $\bar{n}_j$ =100, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
135	d_1	1.0	0.922	-0.078	7.800	0.148
87	d_2	1.0	0.895	-0.105	10.500	0.091
150	d_3	2.0	1.721	-0.279	13.950	0.292
210	d_4	2.0	1.793	-0.207	10.350	0.190
69	d_5	3.0	1.644	-1.356	45.200	2.102
75	d_6	3.0	1.737	-1.263	42.100	1.825
249	d_7	4.0	2.496	-1.504	37.600	2.394
225	d_8	4.0	2.646	-1.354	33.850	1.993
150	d_9	5.0	2.783	-2.217	44.340	5.058
150	d_{10}	5.0	2.667	-2.333	46.660	5.635
135	c_{11}	-1.5	-1.878	-0.378	25.200	0.650
135	c_{12}	-0.5	-0.848	-0.348	69.600	0.503
135	c_{13}	0.5	0.147	-0.353	70.600	0.651
135	c_{14}	1.5	1.169	-0.331	22.070	1.010
135	c_{15}	2.5	2.257	-0.243	9.720	1.427
87	c_{21}	2.5	2.127	-0.373	14.920	1.318
87	c_{22}	3.5	3.173	-0.327	9.340	1.796
87	c_{23}	4.5	4.157	-0.343	7.620	2.435
87	c_{24}	5.5	5.144	-0.356	6.470	3.043
87	c_{25}	6.5	6.206	-0.294	4.520	4.122
150	c_{31}	-1.0	-1.733	-0.733	73.300	0.927
150	c_{32}	1.0	0.065	-0.935	93.500	1.294
150	c_{33}	3.0	2.084	-0.916	30.530	2.005
150	c_{34}	5.0	4.156	-0.844	16.880	2.888
150	c_{35}	7.0	6.272	-0.728	10.400	4.655
210	c_{41}	3.0	2.131	-0.869	28.970	1.780
210	c_{42}	5.0	4.272	-0.728	14.560	2.738
210	c_{43}	7.0	6.480	-0.520	7.430	4.164
210	c_{44}	9.0	8.535	-0.465	5.170	5.727
210	c_{45}	11.0	10.418	-0.582	5.290	7.612
69	c_{51}	-0.5	-1.714	-1.214	242.800	2.789
69	c_{52}	2.5	0.192	-2.308	92.320	6.610
69	c_{53}	5.5	2.441	-3.059	55.620	11.290
69	c_{54}	8.5	4.575	-3.925	46.180	18.503
69	c_{55}	11.5	6.828	-4.672	40.630	27.403
75	c_{61}	3.5	1.164	-2.336	66.740	6.660
75	c_{62}	6.5	3.467	-3.033	46.660	11.743
75	c_{63}	9.5	5.792	-3.708	39.030	17.922
75	c_{64}	12.5	8.015	-4.485	35.880	27.184
75	c_{65}	15.5	10.008	-5.492	35.430	39.930

Table C1.3 (Continued)
30% 2nd rater, $\bar{n}_j$ =100, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
249	c_{71}	0.0	-1.254	-1.254	.	1.819
249	c_{72}	4.0	1.272	-2.728	68.200	8.203
249	c_{73}	8.0	4.444	-3.556	44.450	14.740
249	c_{74}	12.0	7.532	-4.468	37.230	24.289
249	c_{75}	16.0	10.741	-5.259	32.870	35.065
225	c_{81}	4.0	1.345	-2.655	66.380	7.973
225	c_{82}	8.0	4.716	-3.284	41.050	13.491
225	c_{83}	12.0	7.930	-4.070	33.920	22.222
225	c_{84}	16.0	11.337	-4.663	29.140	31.432
225	c_{85}	20.0	13.947	-6.053	30.270	49.701
150	c_{91}	0.5	-0.941	-1.441	288.200	2.473
150	c_{92}	5.5	1.944	-3.556	64.650	14.194
150	c_{93}	10.5	5.410	-5.090	48.480	28.949
150	c_{94}	15.5	8.793	-6.707	43.270	51.128
150	c_{95}	20.5	12.262	-8.238	40.190	77.779
150	c_{101}	4.5	0.914	-3.586	79.690	13.615
150	c_{102}	9.5	4.237	-5.263	55.400	30.017
150	c_{103}	14.5	7.475	-7.025	48.450	54.333
150	c_{104}	19.5	10.839	-8.661	44.420	83.751
150	c_{105}	24.5	13.811	-10.689	43.630	127.181
			Latent Class Sizes			
	Class 1	0.08	0.176	0.096	120.000	
	Class 2	0.17	0.113	-0.057	33.530	
	Class 3	0.25	0.217	-0.033	13.200	
	Class 4	0.25	0.218	-0.032	12.800	
	Class 5	0.17	0.113	-0.057	33.530	
	Class 6	0.08	0.163	0.083	103.750	

Table C1.4
30% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
675	d_1	1.0	0.964	-0.036	3.600	0.029
435	d_2	1.0	0.964	-0.036	3.600	0.019
750	d_3	2.0	1.872	-0.128	6.400	0.092
1050	d_4	2.0	1.907	-0.093	4.650	0.052
345	d_5	3.0	2.532	-0.468	15.600	0.333
375	d_6	3.0	2.537	-0.463	15.433	0.313
1245	d_7	4.0	3.394	-0.606	15.150	0.518
1125	d_8	4.0	3.523	-0.477	11.925	0.423
750	d_9	5.0	4.107	-0.893	17.860	0.964
750	d_{10}	5.0	4.096	-0.904	18.080	0.970
675	c_{11}	-1.5	-1.690	-0.190	12.667	0.112
675	c_{12}	-0.5	-0.684	-0.184	36.800	0.114
675	c_{13}	0.5	0.347	-0.153	30.600	0.123
675	c_{14}	1.5	1.383	-0.117	7.800	0.163
675	c_{15}	2.5	2.401	-0.099	3.960	0.203
435	c_{21}	2.5	2.403	-0.097	3.880	0.143
435	c_{22}	3.5	3.423	-0.077	2.200	0.183
435	c_{23}	4.5	4.433	-0.067	1.489	0.228
435	c_{24}	5.5	5.444	-0.056	1.018	0.283
435	c_{25}	6.5	6.491	-0.009	0.138	0.334
750	c_{31}	-1.0	-1.303	-0.303	30.300	0.204
750	c_{32}	1.0	0.621	-0.379	37.900	0.305
750	c_{33}	3.0	2.614	-0.386	12.867	0.454
750	c_{34}	5.0	4.618	-0.382	7.640	0.675
750	c_{35}	7.0	6.628	-0.372	5.314	0.960
1050	c_{41}	3.0	2.711	-0.289	9.633	0.304
1050	c_{42}	5.0	4.722	-0.278	5.560	0.400
1050	c_{43}	7.0	6.785	-0.215	3.071	0.568
1050	c_{44}	9.0	8.807	-0.193	2.144	0.703
1050	c_{45}	11.0	10.790	-0.210	1.909	0.909
345	c_{51}	-0.5	-0.948	-0.448	89.600	0.447
345	c_{52}	2.5	1.642	-0.858	34.320	1.085
345	c_{53}	5.5	4.409	-1.091	19.836	1.759
345	c_{54}	8.5	7.146	-1.354	15.929	2.796
345	c_{55}	11.5	9.975	-1.525	13.261	4.107
375	c_{61}	3.5	2.601	-0.899	25.686	1.178
375	c_{62}	6.5	5.419	-1.081	16.631	1.797
375	c_{63}	9.5	8.173	-1.327	13.968	2.840
375	c_{64}	12.5	10.964	-1.536	12.288	4.058
375	c_{65}	15.5	13.578	-1.922	12.400	5.907

Table C1.4 (Continued)
30% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
1245	c_{71}	0.0	-0.506	-0.506	.	0.427
1245	c_{72}	4.0	2.922	-1.078	26.950	1.692
1245	c_{73}	8.0	6.617	-1.383	17.288	2.688
1245	c_{74}	12.0	10.234	-1.766	14.717	4.575
1245	c_{75}	16.0	13.905	-2.095	13.094	6.803
1125	c_{81}	4.0	3.033	-0.967	24.175	1.467
1125	c_{82}	8.0	6.890	-1.110	13.875	2.134
1125	c_{83}	12.0	10.615	-1.385	11.542	3.708
1125	c_{84}	16.0	14.419	-1.581	9.881	5.253
1125	c_{85}	20.0	17.952	-2.048	10.240	8.199
750	c_{91}	0.5	-0.096	-0.596	119.200	0.596
750	c_{92}	5.5	4.087	-1.413	25.691	2.726
750	c_{93}	10.5	8.552	-1.948	18.552	4.993
750	c_{94}	15.5	12.878	-2.622	16.916	8.849
750	c_{95}	20.5	17.314	-3.186	15.541	13.200
750	c_{101}	4.5	3.002	-1.498	33.289	2.743
750	c_{102}	9.5	7.523	-1.977	20.811	4.534
750	c_{103}	14.5	11.771	-2.729	18.821	8.701
750	c_{104}	19.5	16.262	-3.238	16.605	12.763
750	c_{105}	24.5	20.448	-4.052	16.539	19.897
	Class 1	0.08	0.108	0.028	35.000	
			Latent Class Sizes			
	Class 2	0.17	0.157	-0.013	7.647	
	Class 3	0.25	0.241	-0.009	3.600	
	Class 4	0.25	0.233	-0.017	6.800	
	Class 5	0.17	0.161	-0.009	5.294	
	Class 6	0.08	0.100	0.020	25.000	

Appendix C2

Evaluation of the Estimated Standard Errors for Bayes' Constants (via PME) in Simulation 2 (Partial Second Rater per Examinee)

Table C2.1
10% 2nd rater, $\bar{n}_j$ =100, PME with Bayes' Constants

Sample size	Parameter	SD	MeanSE	Bias	%Bias
110	d_1	0.469	0.584	0.115	24.520
47	d_2	0.470	0.537	0.068	14.468
110	d_3	0.449	0.660	0.211	46.993
160	d_4	0.423	0.667	0.244	57.683
53	d_5	0.742	0.722	-0.019	2.561
55	d_6	0.748	0.731	-0.017	2.273
185	d_7	0.312	0.745	0.433	138.782
165	d_8	0.344	0.746	0.402	116.860
110	d_9	0.445	0.770	0.324	72.809
110	d_{10}	0.485	0.753	0.268	55.258
	Class size 1	0.055	0.096	0.042	76.364
	Class size 2	0.056	0.115	0.059	105.357
	Class size 3	0.113	0.144	0.032	28.319
	Class size 4	0.114	0.142	0.027	23.684
	Class size 5	0.070	0.121	0.051	72.857
	Class size 6	0.061	0.102	0.041	67.213

Table C2.2
10% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	SD	MeanSE	Bias	%Bias
550	d_1	0.307	0.304	-0.003	0.977
235	d_2	0.247	0.212	-0.035	14.170
550	d_3	0.405	0.429	0.024	5.926
800	d_4	0.316	0.336	0.020	6.329
265	d_5	0.377	0.676	0.298	79.045
275	d_6	0.333	0.696	0.364	109.309
925	d_7	0.415	0.716	0.301	72.530
825	d_8	0.425	0.683	0.258	60.706
550	d_9	0.401	0.791	0.389	97.007
550	d_{10}	0.416	0.804	0.388	93.269
	Class size 1	0.042	0.043	0.001	2.381
	Class size 2	0.056	0.055	-0.001	1.786
	Class size 3	0.059	0.054	-0.005	8.475
	Class size 4	0.054	0.053	-0.001	1.852
	Class size 5	0.051	0.052	0.001	1.961
	Class size 6	0.045	0.041	-0.004	8.889

Table C2.3
30% 2nd rater, $\bar{n}_j$ =100, PME with Bayes' Constants

Sample size	Parameter	SD	MeanSE	Bias	%Bias
135	d_1	0.378	0.390	0.011	2.910
87	d_2	0.284	0.287	0.003	1.056
150	d_3	0.465	0.568	0.103	22.151
210	d_4	0.386	0.463	0.077	19.948
69	d_5	0.516	0.725	0.209	40.504
75	d_6	0.481	0.753	0.272	56.549
249	d_7	0.365	0.720	0.356	97.534
225	d_8	0.402	0.744	0.342	85.075
150	d_9	0.382	0.833	0.451	118.063
150	d_{10}	0.441	0.804	0.363	82.313
	Class size 1	0.040	0.057	0.017	42.500
	Class size 2	0.062	0.073	0.011	17.742
	Class size 3	0.073	0.077	0.004	5.479
	Class size 4	0.074	0.078	0.005	6.757
	Class size 5	0.067	0.069	0.002	2.985
	Class size 6	0.044	0.051	0.007	15.909

Table C2.4
30% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	SD	MeanSE	Bias	%Bias
675	d_1	0.168	0.178	0.009	5.357
435	d_2	0.135	0.134	-0.001	0.741
750	d_3	0.277	0.278	0.001	0.361
1050	d_4	0.209	0.215	0.005	2.392
345	d_5	0.339	0.680	0.341	100.590
375	d_6	0.315	0.687	0.372	118.095
1245	d_7	0.390	0.684	0.294	75.385
1125	d_8	0.444	0.697	0.253	56.982
750	d_9	0.409	0.831	0.421	102.934
750	d_{10}	0.394	0.811	0.417	105.838
	Class size 1	0.029	0.024	-0.004	13.793
	Class size 2	0.029	0.024	-0.005	17.241
	Class size 3	0.027	0.026	-0.001	3.704
	Class size 4	0.025	0.026	0.000	0.000
	Class size 5	0.026	0.022	-0.004	15.385
	Class size 6	0.025	0.022	-0.003	12.000

Appendix D1

Parameter Estimates, Bias, Percent Bias, and MSE for Informative Priors (via MCMC) in Simulation 2 (Partial Second Rater per Examinee)

Table D1.1
10% 2nd rater, $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
110	d_1	1.0	4.201	3.201	320.100	10.976
47	d_2	1.0	1.754	0.754	75.400	0.886
110	d_3	2.0	3.512	1.512	75.600	2.712
160	d_4	2.0	2.997	0.997	49.850	1.197
53	d_5	3.0	4.065	1.065	35.500	1.243
55	d_6	3.0	3.074	0.074	2.470	0.205
185	d_7	4.0	3.925	-0.075	1.880	0.108
165	d_8	4.0	3.465	-0.535	13.380	0.383
110	d_9	5.0	4.040	-0.960	19.200	1.046
110	d_{10}	5.0	3.656	-1.344	26.880	1.896
110	c_{11}	-1.5	-0.242	1.258	83.870	2.199
110	c_{12}	-0.5	1.834	2.334	466.800	6.449
110	c_{13}	0.5	3.992	3.492	698.400	13.674
110	c_{14}	1.5	6.443	4.943	329.530	26.829
110	c_{15}	2.5	9.196	6.696	267.840	48.206
47	c_{21}	2.5	3.438	0.938	37.520	2.003
47	c_{22}	3.5	4.980	1.480	42.290	3.987
47	c_{23}	4.5	6.591	2.091	46.470	7.463
47	c_{24}	5.5	8.245	2.745	49.910	11.554
47	c_{25}	6.5	10.496	3.996	61.480	21.617
110	c_{31}	-1.0	-0.633	0.367	36.700	0.619
110	c_{32}	1.0	1.914	0.914	91.400	1.587
110	c_{33}	3.0	4.663	1.663	55.430	4.134
110	c_{34}	5.0	7.614	2.614	52.280	9.013
110	c_{35}	7.0	10.764	3.764	53.770	17.551
160	c_{41}	3.0	3.617	0.617	20.570	0.769
160	c_{42}	5.0	6.392	1.392	27.840	2.843
160	c_{43}	7.0	9.299	2.299	32.840	6.928
160	c_{44}	9.0	12.289	3.289	36.540	13.243
160	c_{45}	11.0	15.298	4.298	39.070	21.947
53	c_{51}	-0.5	-0.207	0.293	58.600	0.770
53	c_{52}	2.5	2.950	0.450	18.000	1.126
53	c_{53}	5.5	6.457	0.957	17.400	2.012
53	c_{54}	8.5	10.046	1.546	18.190	3.446
53	c_{55}	11.5	13.905	2.405	20.910	6.815

Table D1.1 (Continued)
10% 2nd rater, $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
55	c_{61}	3.5	2.646	-0.854	24.400	1.197
55	c_{62}	6.5	5.699	-0.801	12.320	1.507
55	c_{63}	9.5	8.753	-0.747	7.860	1.832
55	c_{64}	12.5	11.988	-0.512	4.100	2.258
55	c_{65}	15.5	15.492	-0.008	0.050	2.256
185	c_{71}	0.0	-0.163	-0.163	.	0.380
185	c_{72}	4.0	3.171	-0.829	20.730	1.203
185	c_{73}	8.0	6.706	-1.294	16.180	2.485
185	c_{74}	12.0	10.431	-1.569	13.080	3.399
185	c_{75}	16.0	14.391	-1.609	10.060	3.631
165	c_{81}	4.0	2.558	-1.442	36.050	2.308
165	c_{82}	8.0	5.818	-2.182	27.280	5.185
165	c_{83}	12.0	9.198	-2.802	23.350	8.557
165	c_{84}	16.0	12.847	-3.153	19.710	10.963
165	c_{85}	20.0	16.574	-3.426	17.130	12.999
110	c_{91}	0.5	0.338	-0.162	32.400	0.554
110	c_{92}	5.5	3.799	-1.701	30.930	3.641
110	c_{93}	10.5	7.338	-3.162	30.110	10.931
110	c_{94}	15.5	11.151	-4.349	28.060	20.040
110	c_{95}	20.5	15.140	-5.360	26.150	29.976
110	c_{101}	4.5	2.234	-2.266	50.360	5.474
110	c_{102}	9.5	5.612	-3.888	40.930	15.581
110	c_{103}	14.5	9.117	-5.383	37.120	29.823
110	c_{104}	19.5	12.872	-6.628	33.990	44.885
110	c_{105}	24.5	16.945	-7.555	30.840	58.130
	Class 1	0.08	0.101	0.021	26.250	
			Latent Class Sizes			
	Class 2	0.17	0.222	0.052	30.590	
	Class 3	0.25	0.268	0.018	7.200	
	Class 4	0.25	0.224	-0.026	10.400	
	Class 5	0.17	0.105	-0.065	38.240	
	Class 6	0.08	0.079	-0.001	1.250	

Table D1.2
10% 2nd rater, $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
550	d_1	1.0	2.351	1.351	135.100	3.280
235	d_2	1.0	1.196	0.196	19.600	0.128
550	d_3	2.0	2.816	0.816	40.800	1.370
800	d_4	2.0	2.486	0.486	24.300	0.394
265	d_5	3.0	4.058	1.058	35.270	1.229
275	d_6	3.0	2.973	-0.027	0.900	0.209
925	d_7	4.0	4.135	0.135	3.380	0.158
825	d_8	4.0	3.603	-0.397	9.930	0.313
550	d_9	5.0	4.128	-0.872	17.440	0.853
550	d_{10}	5.0	3.967	-1.033	20.660	1.142
550	c_{11}	-1.5	-0.900	0.600	40.000	0.931
550	c_{12}	-0.5	0.576	1.076	215.200	2.494
550	c_{13}	0.5	2.117	1.617	323.400	5.196
550	c_{14}	1.5	3.849	2.349	156.600	10.717
550	c_{15}	2.5	5.538	3.038	121.520	17.605
235	c_{21}	2.5	2.804	0.304	12.160	0.657
235	c_{22}	3.5	3.903	0.403	11.510	0.977
235	c_{23}	4.5	4.978	0.478	10.620	1.265
235	c_{24}	5.5	6.101	0.601	10.930	1.678
235	c_{25}	6.5	7.359	0.859	13.220	2.392
550	c_{31}	-1.0	-0.770	0.230	23.000	0.417
550	c_{32}	1.0	1.579	0.579	57.900	1.123
550	c_{33}	3.0	4.015	1.015	33.830	2.936
550	c_{34}	5.0	6.548	1.548	30.960	6.242
550	c_{35}	7.0	9.100	2.100	30.000	11.058
800	c_{41}	3.0	3.544	0.544	18.130	0.738
800	c_{42}	5.0	5.812	0.812	16.240	1.531
800	c_{43}	7.0	8.133	1.133	16.190	2.787
800	c_{44}	9.0	10.539	1.539	17.100	4.646
800	c_{45}	11.0	12.948	1.948	17.710	6.646
265	c_{51}	-0.5	-0.029	0.471	94.200	0.829
265	c_{52}	2.5	3.372	0.872	34.880	1.412
265	c_{53}	5.5	7.007	1.507	27.400	3.134
265	c_{54}	8.5	10.710	2.210	26.000	5.943
265	c_{55}	11.5	14.479	2.979	25.900	10.330
275	c_{61}	3.5	3.101	-0.399	11.400	0.591
275	c_{62}	6.5	6.015	-0.485	7.460	1.127
275	c_{63}	9.5	8.882	-0.618	6.510	2.120
275	c_{64}	12.5	11.882	-0.618	4.940	3.249
275	c_{65}	15.5	14.952	-0.548	3.540	4.110

Table D1.2 (Continued)
10% 2nd rater, $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
925	c_{71}	0.0	0.221	0.221	.	0.479
925	c_{72}	4.0	3.928	-0.072	1.800	0.689
925	c_{73}	8.0	7.777	-0.223	2.790	1.048
925	c_{74}	12.0	11.648	-0.352	2.930	1.522
925	c_{75}	16.0	15.613	-0.387	2.420	1.989
825	c_{81}	4.0	3.311	-0.689	17.230	1.028
825	c_{82}	8.0	6.753	-1.247	15.590	2.518
825	c_{83}	12.0	10.187	-1.813	15.110	4.791
825	c_{84}	16.0	13.718	-2.282	14.260	7.446
825	c_{85}	20.0	17.382	-2.618	13.090	9.760
550	c_{91}	0.5	0.604	0.104	20.800	0.494
550	c_{92}	5.5	4.333	-1.167	21.220	1.992
550	c_{93}	10.5	8.239	-2.261	21.530	5.987
550	c_{94}	15.5	12.149	-3.351	21.620	12.520
550	c_{95}	20.5	16.216	-4.284	20.900	20.070
550	c_{101}	4.5	3.171	-1.329	29.530	2.255
550	c_{102}	9.5	6.951	-2.549	26.830	7.198
550	c_{103}	14.5	10.683	-3.817	26.320	15.521
550	c_{104}	19.5	14.612	-4.888	25.070	25.368
550	c_{105}	24.5	18.675	-5.825	23.780	35.321
			Latent Class Sizes			
	Class 1	0.08	0.084	0.004	5.000	
	Class 2	0.17	0.192	0.022	12.940	
	Class 3	0.25	0.268	0.018	7.200	
	Class 4	0.25	0.247	-0.003	1.200	
	Class 5	0.17	0.137	-0.033	19.410	
	Class 6	0.08	0.072	-0.008	10.000	

Table D1.3
30% 2nd rater, $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
135	d_1	1.0	2.778	1.778	177.800	4.688
87	d_2	1.0	1.277	0.277	27.700	0.198
150	d_3	2.0	3.135	1.135	56.750	2.000
210	d_4	2.0	2.649	0.649	32.450	0.598
69	d_5	3.0	4.098	1.098	36.600	1.349
75	d_6	3.0	3.087	0.087	2.900	0.216
249	d_7	4.0	4.015	0.015	0.380	0.105
225	d_8	4.0	3.542	-0.458	11.450	0.356
150	d_9	5.0	4.127	-0.873	17.460	0.872
150	d_{10}	5.0	3.802	-1.198	23.960	1.552
135	c_{11}	-1.5	-0.750	0.750	50.000	1.489
135	c_{12}	-0.5	0.929	1.429	285.800	3.749
135	c_{13}	0.5	2.611	2.111	422.200	7.698
135	c_{14}	1.5	4.429	2.929	195.270	14.425
135	c_{15}	2.5	6.379	3.879	155.160	24.156
87	c_{21}	2.5	2.762	0.262	10.480	0.790
87	c_{22}	3.5	3.956	0.456	13.030	1.191
87	c_{23}	4.5	5.116	0.616	13.690	1.684
87	c_{24}	5.5	6.385	0.885	16.090	2.397
87	c_{25}	6.5	8.103	1.603	24.660	5.294
150	c_{31}	-1.0	-0.731	0.269	26.900	0.486
150	c_{32}	1.0	1.696	0.696	69.600	1.272
150	c_{33}	3.0	4.299	1.299	43.300	3.511
150	c_{34}	5.0	6.988	1.988	39.760	7.124
150	c_{35}	7.0	9.769	2.769	39.560	13.300
210	c_{41}	3.0	3.397	0.397	13.230	0.605
210	c_{42}	5.0	5.866	0.866	17.320	1.606
210	c_{43}	7.0	8.401	1.401	20.010	3.321
210	c_{44}	9.0	10.972	1.972	21.910	6.011
210	c_{45}	11.0	13.648	2.648	24.070	9.721
69	c_{51}	-0.5	-0.142	0.358	71.600	0.822
69	c_{52}	2.5	3.088	0.588	23.520	1.270
69	c_{53}	5.5	6.704	1.204	21.890	2.494
69	c_{54}	8.5	10.277	1.777	20.910	4.126
69	c_{55}	11.5	14.081	2.581	22.440	7.623
75	c_{61}	3.5	2.881	-0.619	17.690	0.871
75	c_{62}	6.5	5.947	-0.553	8.510	1.147
75	c_{63}	9.5	8.964	-0.536	5.640	1.624
75	c_{64}	12.5	12.053	-0.447	3.580	2.225
75	c_{65}	15.5	15.390	-0.110	0.710	2.542

Table D1.3 (Continued)
30% 2nd rater, $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
249	c_{71}	0.0	-0.093	-0.093		0.338
249	c_{72}	4.0	3.435	-0.565	14.130	0.816
249	c_{73}	8.0	7.128	-0.872	10.900	1.622
249	c_{74}	12.0	10.917	-1.083	9.030	2.211
249	c_{75}	16.0	14.830	-1.170	7.310	2.652
225	c_{81}	4.0	2.834	-1.166	29.150	1.659
225	c_{82}	8.0	6.211	-1.789	22.360	3.779
225	c_{83}	12.0	9.610	-2.390	19.920	6.720
225	c_{84}	16.0	13.214	-2.786	17.410	9.325
225	c_{85}	20.0	16.875	-3.125	15.630	11.822
150	c_{91}	0.5	0.379	-0.121	24.200	0.510
150	c_{92}	5.5	4.106	-1.394	25.350	2.672
150	c_{93}	10.5	7.836	-2.664	25.370	8.005
150	c_{94}	15.5	11.722	-3.778	24.370	15.525
150	c_{95}	20.5	15.657	-4.843	23.620	24.839
150	c_{101}	4.5	2.556	-1.944	43.200	4.177
150	c_{102}	9.5	6.124	-3.376	35.540	11.936
150	c_{103}	14.5	9.752	-4.748	32.740	23.266
150	c_{104}	19.5	13.483	-6.017	30.860	37.390
150	c_{105}	24.5	17.565	-6.935	28.310	49.557
			Latent Class Sizes			
	Class 1	0.08	0.098	0.018	22.500	
	Class 2	0.17	0.204	0.034	20.000	
	Class 3	0.25	0.271	0.021	8.400	
	Class 4	0.25	0.236	-0.014	5.600	
	Class 5	0.17	0.117	-0.053	31.180	
	Class 6	0.08	0.075	-0.005	6.250	

Table D1.4
30% 2nd rater, $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
675	d_1	1.0	1.203	0.203	20.300	0.138
435	d_2	1.0	1.066	0.066	6.600	0.026
750	d_3	2.0	2.347	0.347	17.350	0.300
1050	d_4	2.0	2.164	0.164	8.200	0.071
345	d_5	3.0	3.811	0.811	27.030	0.773
375	d_6	3.0	3.022	0.022	0.730	0.120
1245	d_7	4.0	4.089	0.089	2.230	0.111
1125	d_8	4.0	3.723	-0.277	6.930	0.191
750	d_9	5.0	4.358	-0.642	12.840	0.497
750	d_{10}	5.0	4.270	-0.730	14.600	0.588
675	c_{11}	-1.5	-1.408	0.092	6.130	0.141
675	c_{12}	-0.5	-0.335	0.165	33.000	0.212
675	c_{13}	0.5	0.769	0.269	53.800	0.344
675	c_{14}	1.5	1.885	0.385	25.670	0.583
675	c_{15}	2.5	2.984	0.484	19.360	0.844
435	c_{21}	2.5	2.623	0.123	4.920	0.161
435	c_{22}	3.5	3.664	0.164	4.690	0.220
435	c_{23}	4.5	4.700	0.200	4.440	0.279
435	c_{24}	5.5	5.754	0.254	4.620	0.362
435	c_{25}	6.5	6.885	0.385	5.920	0.506
750	c_{31}	-1.0	-0.856	0.144	14.400	0.183
750	c_{32}	1.0	1.295	0.295	29.500	0.384
750	c_{33}	3.0	3.507	0.507	16.900	0.888
750	c_{34}	5.0	5.736	0.736	14.720	1.707
750	c_{35}	7.0	7.988	0.988	14.110	2.854
1050	c_{41}	3.0	3.226	0.226	7.530	0.266
1050	c_{42}	5.0	5.316	0.316	6.320	0.452
1050	c_{43}	7.0	7.464	0.464	6.630	0.782
1050	c_{44}	9.0	9.593	0.593	6.590	1.102
1050	c_{45}	11.0	11.737	0.737	6.700	1.528
345	c_{51}	-0.5	-0.108	0.392	78.400	0.700
345	c_{52}	2.5	3.285	0.785	31.400	1.187
345	c_{53}	5.5	6.905	1.405	25.550	2.688
345	c_{54}	8.5	10.554	2.054	24.160	5.217
345	c_{55}	11.5	14.267	2.767	24.060	9.168
375	c_{61}	3.5	3.398	-0.102	2.910	0.369
375	c_{62}	6.5	6.437	-0.063	0.970	0.646
375	c_{63}	9.5	9.437	-0.063	0.660	1.140
375	c_{64}	12.5	12.471	-0.029	0.230	1.759
375	c_{65}	15.5	15.484	-0.016	0.100	2.275

Table C1.4 (Continued)
30% 2nd rater, $\bar{n}_j$ =500, PME with Bayes' Constants

Sample size	Parameter	Value	Estimate	Bias	%Bias	MSE
1245	c_{71}	0.0	0.202	0.202	.	0.380
1245	c_{72}	4.0	4.054	0.054	1.350	0.483
1245	c_{73}	8.0	8.062	0.062	0.780	0.700
1245	c_{74}	12.0	12.110	0.110	0.920	1.174
1245	c_{75}	16.0	16.148	0.148	0.930	1.844
1125	c_{81}	4.0	3.601	-0.399	9.980	0.504
1125	c_{82}	8.0	7.322	-0.678	8.480	1.036
1125	c_{83}	12.0	11.020	-0.980	8.170	2.027
1125	c_{84}	16.0	14.736	-1.264	7.900	3.236
1125	c_{85}	20.0	18.378	-1.622	8.110	4.857
750	c_{91}	0.5	0.680	0.180	36.000	0.476
750	c_{92}	5.5	4.781	-0.719	13.070	1.026
750	c_{93}	10.5	9.089	-1.411	13.440	2.768
750	c_{94}	15.5	13.397	-2.103	13.570	5.580
750	c_{95}	20.5	17.739	-2.761	13.470	9.295
750	c_{101}	4.5	3.608	-0.892	19.820	1.083
750	c_{102}	9.5	7.894	-1.606	16.910	2.850
750	c_{103}	14.5	12.071	-2.429	16.750	6.359
750	c_{104}	19.5	16.382	-3.118	15.990	10.463
750	c_{105}	24.5	20.563	-3.937	16.070	16.566
			Latent Class Sizes			
	Class 1	0.08	0.081	0.001	1.250	
	Class 2	0.17	0.174	0.004	2.350	
	Class 3	0.25	0.256	0.006	2.400	
	Class 4	0.25	0.249	-0.001	0.400	
	Class 5	0.17	0.164	-0.006	3.530	
	Class 6	0.08	0.076	-0.004	5.000	

Appendix D2

Evaluation of the Estimated Posterior Standard Deviations for Informative Priors (via MCMC) in Simulation 2 (Partial Second Rater per Examinee)

Table D2.1
10% 2nd rater, $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
110	d_1	0.857	1.419	0.562	65.578
47	d_2	0.567	0.628	0.061	10.758
110	d_3	0.655	1.096	0.440	67.176
160	d_4	0.454	0.717	0.263	57.930
53	d_5	0.331	1.025	0.694	209.668
55	d_6	0.449	0.845	0.396	88.196
185	d_7	0.322	0.933	0.611	189.752
165	d_8	0.313	0.771	0.458	146.326
110	d_9	0.355	0.906	0.552	155.493
110	d_{10}	0.301	0.815	0.515	171.096
	Class size 1	0.021	0.038	0.017	80.952
	Class size 2	0.039	0.057	0.018	46.154
	Class size 3	0.036	0.063	0.028	77.778
	Class size 4	0.039	0.061	0.022	56.410
	Class size 5	0.027	0.055	0.028	103.704
	Class size 6	0.018	0.046	0.028	155.556

Table D2.2
10% 2nd rater, $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
550	d_1	1.212	1.045	-0.167	13.779
235	d_2	0.301	0.274	-0.027	8.970
550	d_3	0.843	0.821	-0.022	2.610
800	d_4	0.398	0.449	0.050	12.563
265	d_5	0.332	0.920	0.588	177.108
275	d_6	0.459	0.734	0.275	59.913
925	d_7	0.376	0.772	0.396	105.319
825	d_8	0.397	0.627	0.230	57.935
550	d_9	0.307	0.767	0.459	149.511
550	d_{10}	0.277	0.682	0.404	145.848
	Class size 1	0.027	0.029	0.002	7.407
	Class size 2	0.035	0.036	0.001	2.857
	Class size 3	0.036	0.042	0.006	16.667
	Class size 4	0.033	0.039	0.005	15.152
	Class size 5	0.035	0.038	0.003	8.571
	Class size 6	0.018	0.036	0.018	100.000

Table D2.3
30% 2nd rater, $\bar{n}_j$ =100, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
135	d_1	1.242	1.243	0.001	0.081
87	d_2	0.350	0.369	0.019	5.429
150	d_3	0.847	0.994	0.147	17.355
210	d_4	0.423	0.582	0.159	37.589
69	d_5	0.382	0.984	0.602	157.592
75	d_6	0.459	0.811	0.353	76.906
249	d_7	0.326	0.850	0.524	160.736
225	d_8	0.385	0.711	0.326	84.675
150	d_9	0.332	0.837	0.505	152.108
150	d_{10}	0.343	0.746	0.403	117.493
	Class size 1	0.022	0.035	0.013	59.091
	Class size 2	0.038	0.050	0.012	31.579
	Class size 3	0.035	0.054	0.019	54.286
	Class size 4	0.032	0.053	0.020	62.500
	Class size 5	0.034	0.048	0.014	41.176
	Class size 6	0.021	0.040	0.019	90.476

Table D2.4
30% 2nd rater, $\bar{n}_j$ =500, MCMC with Informative Priors

Sample size	Parameter	SD	MeanSD	Bias	%Bias
675	d_1	0.313	0.277	-0.036	11.502
435	d_2	0.146	0.142	-0.004	2.740
750	d_3	0.427	0.414	-0.013	3.044
1050	d_4	0.212	0.234	0.022	10.377
345	d_5	0.341	0.814	0.473	138.710
375	d_6	0.348	0.610	0.262	75.287
1245	d_7	0.322	0.634	0.312	96.894
1125	d_8	0.340	0.502	0.162	47.647
750	d_9	0.294	0.663	0.369	125.510
750	d_{10}	0.237	0.576	0.339	143.038
	Class size 1	0.022	0.021	-0.001	4.545
	Class size 2	0.021	0.022	0.001	4.762
	Class size 3	0.021	0.024	0.003	14.286
	Class size 4	0.022	0.023	0.002	9.091
	Class size 5	0.021	0.022	0.001	4.762
	Class size 6	0.016	0.022	0.007	43.750

Appendix E

Evaluation of the Convergence in MCMC in Simulation (An Example)

Table E.1

Parameter Estimates in Replication 1, Constant Detection (d=2), $\bar{n}_j$ =100

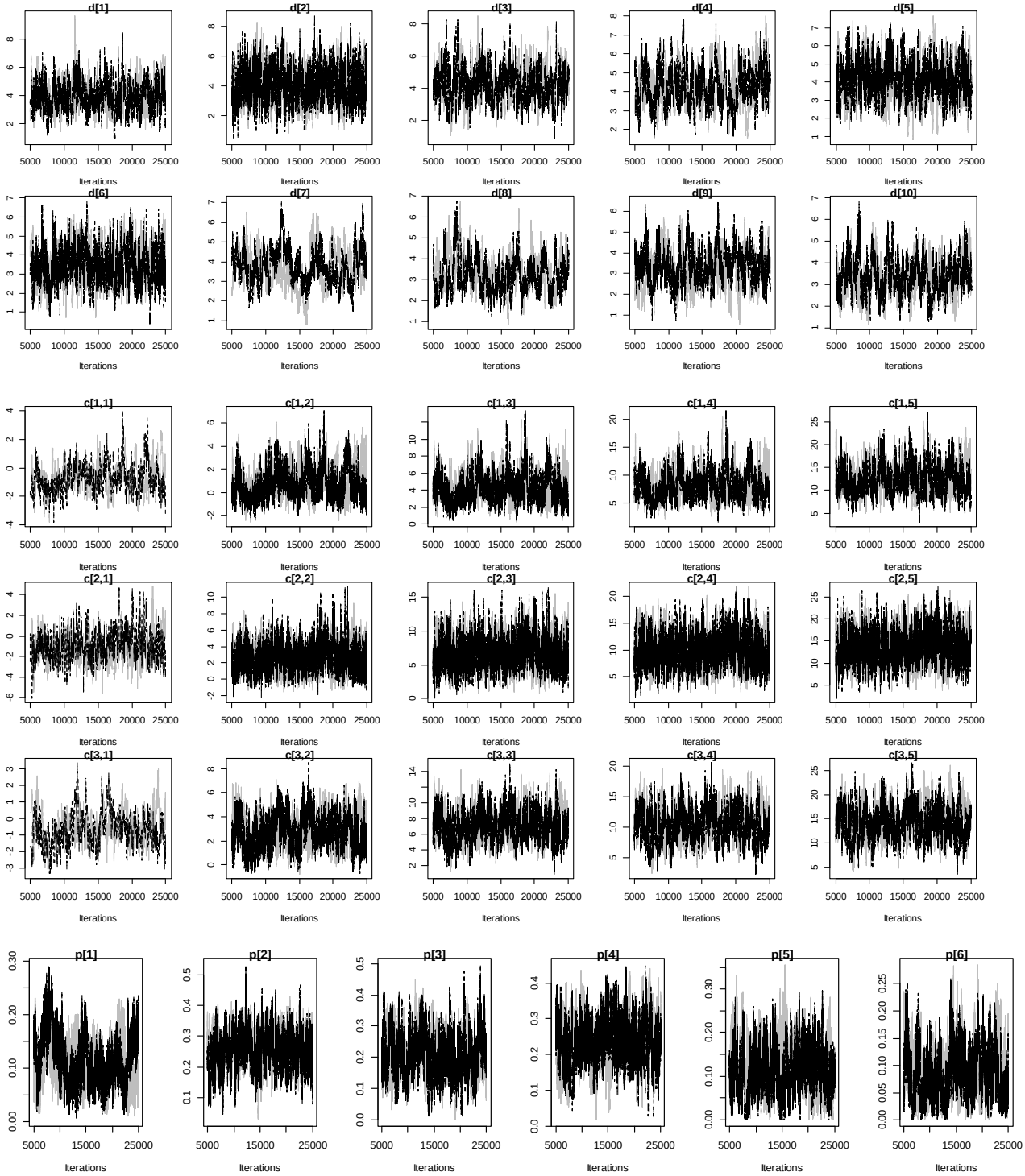
Parameter	Mean	SD	MC error
d_1	3.935	0.964	0.043
d_2	4.003	1.055	0.035
d_3	4.217	1.015	0.046
d_4	4.207	0.955	0.047
d_5	4.070	0.967	0.041
d_6	3.455	0.879	0.037
d_7	3.726	0.907	0.049
d_8	3.288	0.813	0.043
d_9	3.238	0.834	0.042
d_{10}	3.336	0.799	0.041
c_{11}	-0.683	0.967	0.049
c_{12}	0.761	1.233	0.058
c_{13}	4.327	1.579	0.072
c_{14}	8.273	2.370	0.108
c_{15}	12.600	3.018	0.135
c_{21}	-0.971	1.384	0.060
c_{22}	2.681	1.578	0.059
c_{23}	6.762	2.159	0.074
c_{24}	9.928	2.764	0.094
c_{25}	13.500	3.315	0.109
c_{31}	-0.585	0.981	0.051
c_{32}	2.837	1.277	0.058
c_{33}	7.136	1.898	0.081
c_{34}	10.340	2.584	0.114
c_{35}	14.070	3.198	0.141
c_{41}	0.768	1.018	0.056
c_{42}	3.531	1.259	0.063
c_{43}	6.464	1.855	0.089
c_{44}	10.970	2.610	0.127
c_{45}	14.800	3.168	0.152
c_{51}	1.801	1.216	0.059
c_{52}	5.221	1.683	0.071
c_{53}	8.806	2.336	0.098
c_{54}	12.130	2.891	0.121
c_{55}	16.330	3.493	0.142

Table E.1 (Continued)

Parameter Estimates in Replication 1, Constant Detection (d=2), $\bar{n}_j$ =100

Parameter	Mean	SD	MC error
c_{61}	0.447	1.161	0.057
c_{62}	4.042	1.553	0.067
c_{63}	7.478	2.251	0.094
c_{64}	11.640	2.913	0.120
c_{65}	15.830	3.552	0.143
c_{71}	2.869	1.062	0.060
c_{72}	6.832	1.757	0.094
c_{73}	10.220	2.399	0.128
c_{74}	13.500	3.004	0.158
c_{75}	16.760	3.601	0.187
c_{81}	2.782	0.910	0.049
c_{82}	5.378	1.414	0.073
c_{83}	8.529	2.084	0.107
c_{84}	12.720	2.823	0.143
c_{85}	16.230	3.421	0.172
c_{91}	2.544	1.048	0.056
c_{92}	6.055	1.709	0.085
c_{93}	9.845	2.440	0.121
c_{94}	13.750	3.172	0.154
c_{95}	16.590	3.614	0.174
c_{101}	3.665	1.118	0.061
c_{102}	6.627	1.713	0.087
c_{103}	9.991	2.303	0.115
c_{104}	13.670	2.933	0.143
c_{105}	16.960	3.411	0.165
	Latent Class Sizes		
Class 1	0.109	0.044	0.002
Class 2	0.254	0.057	0.003
Class 3	0.204	0.065	0.003
Class 4	0.235	0.059	0.003
Class 5	0.115	0.055	0.003
Class 6	0.084	0.046	0.002

Figure E.1. Trace Plots in Replication 1, Constant Detection (d=2), $\overline{n}_j$ =100



Appendix F

Evaluation of the Convergence in MCMC in Empirical Study

Figure F.1. Trace Plots for Item 4

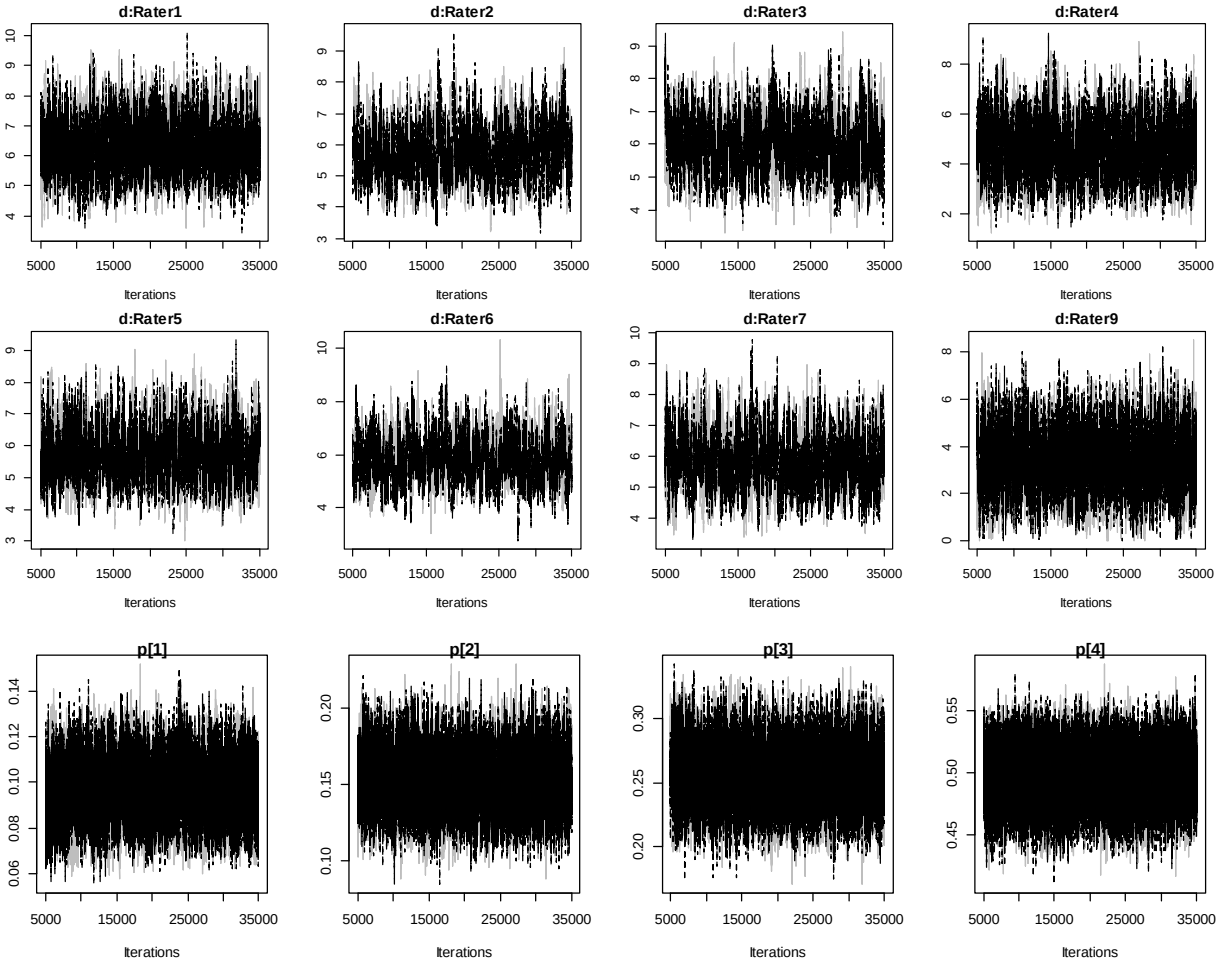


Figure F.2 Trace Plots for Item 5